ERROR COMES WITH IMAGINATION: A PROBABILISTIC THEORY OF MENTAL CONTENT

Hilmi M. Demir

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Departments of Philosophy and Cognitive Science
Indiana University
August 2006

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Frederick Schmitt - Chair, Ph.D.

Colin Allen - Co-Chair, Ph.D.

Timothy O'Connor, Ph.D.

Jonathan Weinberg, Ph.D.

15 August 2006

Dedicated To My Teachers ...

'I would be a slave of those who would teach me even a letter; I would be a slave for them for 40 years.'

Hazreti Ali

Acknowledgements

Whereof one cannot speak
thereof one must be silent.

The last Tractarian Proposition - Ludwig Wittgenstein

Abstract

In this dissertation, I develop a probabilistic theory of mental content that accounts for fundamental properties of mental representation. The theory that I develop falls under the category of causal/informational approaches. In contemporary philosophy, causal/informational approaches for explaining mental representation have been around since the 1950s. The main success of these approaches is to explain the link between the external world and mental entities. On the other hand, it turns out that accounting for misrepresentation cases presents an insurmountable problem for these approaches. The probabilistic theory that I develop keeps the positive features of causal/informational approaches and provides grounds for solving the problem of misrepresentation. The theory that I offer heavily draws from Dretske's 1981 framework. His framework borrows some notions such as entropy from Shannon & Weaver's *Mathematical Theory of Communication* which is a very successful formalism for engineering purposes. Dretske tries to explain mental representation, belief and knowledge by using the notion of informational content. Despite all of its appeals, however, the problem of misrepresentation also afflicts his framework. In this dissertation, I identify the reasons that cause problems in Dretske's framework. Identifying these reasons provides enough grounds for solving the problem of misrepresentation in the theory that I construct. I claim that the theory that I offer not only solves the problem of misrepresentation but also provides a research program for Cognitive Science and Neuroscience.

Contents

Introduction

Seldom do more than a few of nature's secrets give way at one time.

Claude Shannon

In this dissertation, I develop a probabilistic theory of mental content that accounts for fundamental properties of mental representation. The theory that I develop falls under the category of causal/informational approaches. In contemporary philosophy, causal/informational approaches for explaining mental representation have been around since the 1950s. The main success of these approaches is to explain the link between the external world and mental entities. On the other hand, it turns out that accounting for misrepresentation cases presents an insurmountable problem for these approaches. The probabilistic theory that I develop keeps the positive features of causal/informational approaches and provides grounds for solving the problem of misrepresentation. The theory that I offer heavily draws from Dretske's 1981 framework. His framework borrows some notions such as entropy from Shannon & Weaver's *Mathematical Theory of Communication* which is a very successful formalism for engineering purposes. Dretske tries to explain mental representation, belief and knowledge by using the notion of informational content. Despite all of its appeals, however, the problem of misrepresentation also afflicts his framework. In this dissertation, I identify the reasons that cause problems in Dretske's framework. Identifying these reasons provides enough grounds for solving the problem of misrepresentation in the theory that I construct. I claim that the theory that I offer not only solves the problem of misrepresentation but also provides a research program for Cognitive Science and Neuroscience.

The organizational and argumentative structure of this dissertation is as follows. In Chapter 1, I provide a brief survey of the contemporary theories of mental content. None of these theories successfully accounts for the fundamental properties of mental content

and mental representation. I briefly discuss the reasons for why that is the case. Such a discussion leads to five preliminaries that a successful theory of mental content needs to satisfy. These preliminaries provide justification for choosing a causal/informational approach. Dretske's 1981 framework seems to be the best candidate among all other causal/informational approaches because it satisfies four of the five preliminaries. The preliminary requirement that it fails to satisfy is related to the problem of misrepresentation. Dretske attempted to solve the problem of misrepresentation. Chapter 2 discusses his attempt and the shortcomings of his solution. In Chapter 3, I discuss two different perspectives by which mental entities can be studied: the first person perspective and the third person perspective. I analyze the advantages of the first person perspective over the third person perspective. Dretske's framework is the first attempt to use the former in the philosophical literature. However, because of one fundamental feature of his notion of informational content (i.e. assigning unity to conditional probabilities in the definition), he reverses his use of the first person's perspective and ends up with the most extreme version of the third person's perspective (the ideal observer perspective). In Chapter 4, I survey Dretske's reasons for assigning unity to conditional probabilities, and offer arguments against his reasons. In addition to that, I show the plausibility of using Dretske's framework without assigning unity to conditional probabilities and state the first steps of defining informational and mental content in this manner. Several attempts have been made to avoid Dretske's strict constraint by using the notion of mutual information. In Chapter 5, I analyze some of those attempts and argue against using the notion of mutual of information, and defend the claim that Dretske's strict constraint needs to be avoided by using conditional probabilities not by mutual information. In Chapter 6, I provide the fundamentals of the theory that I offer together with the solution to the problem of misrepresentation. Then, I discuss some possible objections to the solution that I offer and refute them.

CHAPTER 1

Contemporary Theories of Mental Representation

Our mental states are very powerful. They give rise to other mental states, for example my belief that Berivan is in town might lead to another belief such as Berivan will call me up tonight. They also give me more or less reliable information regarding the external world. If my mental state implied that a lion is approaching me, my nervous system would start to produce adrenalin and I'd rush to find a place to hide. The main question is how our mental states get their content by which they affect other mental states and physical happenings in the human body as well as observable actions. Mental states get their power from two main dimensions: their connection with the external world and their ability to imply (or cause) other mental states. The former dimension is generally explained by the causal effects of external states of affairs on our mental apparatus. The idea is very simple: our mental states are caused by the external world, and therefore they give us more or less reliable information. For the latter dimension, inferential role is the key. One mental state, because of its specific content, has some epistemic liaisons (borrowing Fodor's coinage) and these epistemic liaisons explain why some of my mental states lead me to have some others. In the rest of the chapter, I will use 'referential power' for the former dimension, and 'inferential power' for the latter one[1].

Contemporary theories of mental content can be categorized with respect to the dimension they prioritize. Some people like Fodor and Dretske put more emphasis on the referential power whereas some others like Harman (1982) and Loar (1981) put more emphasis on the inferential dimension. Theories that are in line with Fodor's and Dretske's are called causal theories of mental content. Conceptual role theories are the ones that belong

[1]It is only natural to ask why not focus on the power to produce actions instead of focusing on the power of leading to other mental states. As I will argue in the later chapters, attempts to use features like function and goal are not able to give us a naturalistic theory of mental content.

to the other party. The above analysis obviously implies there is a third option: putting equal emphasis on both dimensions. Such an approach is the basis of what is commonly known as Two-factor theories [2].

In this chapter, I present a survey of these three different types and analyze the problems that each of these types have. None of these theories gives us a satisfactory theory of mental content because of the problems that they bring about. Analyzing the problems associated with these different theories provides an opportunity for identifying the features that a satisfactory theory of mental content should have. I identify five main features in the last section and call these features preliminaries for a successful theory of mental content. Lastly, I briefly discuss Dretske's theory and show that his 1981 framework satisfies four out of the five preliminaries, and hence provides a good starting point for constructing a new and successful theory of mental content.

1. Causal Theories

The main claim of causal theories is that mental representations acquire their content via what causes them. The mental representation that I have when I see a table is caused by the existence of a table in my visual field, and this is the case for all five sensory modalities that human beings have. Such a view gives room for identifying the truth conditions of mental states. The mental entity TABLE [3] as a representation is triggered when there exists a table in my visual field. In other words, the existence of a table is sufficient for having the representation of 'table'. Obviously, it is not necessary since one can have such a representation as an effect of other mental states, for instance the feeling of guilt that comes with an unfinished paper might lead to thoughts about your study table.

Another motivation for a causal approach is naturalism. A completely physical explanation of mental states has been a goal for philosophers since the time of early physicists of Ancient Greece (reference needed). This goal has become the central tenet in philosophy

[2]For such a categorization please see [Crumley 1999]

[3]Following the convention in the literature, I use capitalized words for mental representations, for example 'DOG', and non-capitalized ones for entities in the external world, for example 'a dog'.

of mind and epistemology after Quine's notion of naturalized epistemology. Causal effects of external states of affairs are prone to scientific explanation, and the causal - scientific explanation avoids any kind of 'mysterious' entity that might enter into our understanding of mental states. Many philosophers followed that path in the past. Some of these are Aristotle, The Stoics, John Locke and Bertrand Russell.

The contemporary discussion of mental representation brought another support for causal theories of mental content. Putnam's and Burge's 'Twin Earth' thought experiments led to a more or less common agreement regarding the importance of external states of affairs for a theory of mental content and meaning. In a sense, externalism has become an indispensable part of our philosophical discourse since then. Fodor's and Dretske's theories are two main influential examples of causal theories of content[4].

Fodor, as other proponents of causal theories, is a realist about propositional attitudes. The network of mental states resembles the network of the propositions that are implied by these mental states, and the relation between these two is mediated by a third entity: mental representation. Hence, his approach regarding propositional attitudes is not monadic, to wit, to have a belief or a desire is to be related in a certain way to a mental representation. In his own formulation,

> O (organism) believes that P (proposition) if and only if O bears R (the believing relation) to M (mental representation) [Fodor 1994]

Mental representations can be simple or complex. Simple ones constitute the primitives of the mentalese that is nothing but a language of thought. The contents of these are fixed by the causal effects of external states of affairs. With this claim, Fodor takes care of the referential power of mental representations. It is important to note that causal relations that fix the content of primitive representations have to be synchronic, that is to say the history of an organism in terms of its causal interactions with the external world does not enter as

[4]One could argue that externalism does not necessitate a causal theory. However, the available externalist theories introduce a causal explanation of one form other in theories. Putnam's and Burge's theories are two prominent examples. Without giving much justification, I assume that externalism requires a causal explanation.

an essential factor into the process of determining the content of a primitive representation. It all happens on the spot.[5]

In Fodor's theory, the content of a complex representation is a function of its meaningful constituents and its computational properties. These computational properties, which stem from the syntactical structure of mentalese, get interpreted by assigning 'meaning' to each of the semantical primitives of the language. With these 'meaning' assignments computational states become cognitive states. The process here is pretty similar to the model theoretic accounts of logic. You have your formulas that are recursively formed out of primitives (individuals, predicates etc.) This is the realm of syntax. Only when some elements, for example numbers, are attached to primitives do you proceed to the semantical realm. In short, language of thought hypothesis requires a strict correlation between syntax and semantics (or between form and content). By introducing the computational properties of mental states, Fodor aims to handle the inferential power of mental representations.

The main problem, besides several others, that Fodor's theory (the causal part of the theory for primitive representations) suffers from is the problem of misrepresentation. It simply goes as following. There is a law-like relationship between my mental representation DOG and dogs. However, the same mental representation can be caused by another object[6], for instance by a cat. The content of DOG must be fixed, to wit, it should refer only to dogs. However, on the other hand, if the content of a mental representation is fixed by the entities that cause them (which is the main claim of causal theories) then it would not be possible to explain cases where the mental representation DOG is triggered by something other than a dog. It is more than likely that in some cases my mental representation DOG is triggered, for example, by a cat. In fact, Fodor raised this objection against the Wisconsin Semantics, and nicely dubbed it as the disjunction problem. Although he raised this objection against other theories, it holds ground for his theory as well.

[5]Fodor's idea of causal interactions happening on the spot is the fundamental motivation behind the synchronic level of informational content that I use in my theory. Please see Chapters 5 and 7 for details.

[6]The term object is used for both animate and inanimate entities.

Fodor's way of fixing this problem in his theory is well known: the asymmetrical causal dependence. The nomic relation between dogs and DOG does not rule out that some other objects may cause DOG as a mental representation. Thus, there are two possible ways, in this simplified example, of having the mental representation, DOG.

R_1: Dogs cause DOG.

R_2: Cats cause DOG

Fodor claims that these two causal relations are radically different in terms of their nomic status. R_1 does not depend on any other nomic relation. It comes into existence and persists by itself. On the other hand, R_2 owes its existence to the former one; it depends on the nomic relation between dogs and DOG. In other words, had there not been R_1, R_2 would not have existed. The name of his solution should be clear after this brief explanation. Dependency: R_2 depends on R_1, Asymmetrical: R_1 does not depend on any other relation, Causal: the nature of the relation is causal, and hence, the asymmetrical causal dependency.

The main problem with Fodor's solution, as pointed out by Cummins [1996], is that it does not solve the problem; rather it hides it behind the veil of dependency. The question becomes what determines the asymmetrical dependency relation between R_1 and R_2. There is no way of judging the merit of his solution unless Fodor gives a precise answer for that question. Unfortunately, Fodor does not provide such an answer. It is not unfair to say that Fodor's asymmetrical dependency does nothing but replace the riddle of misrepresentation by the riddle of asymmetric dependence. To put it in different words, it is true that R1 and R2 have different nomic statuses with respect to the content of mental representation that they trigger. However, this is exactly what the problem of misrepresentation is. Why do R1 and R2 have different nomic statuses? Just to say that they are nomically different is nothing but begging the question. The charge here is the most basic fallacy: petitio principii. The reason why Fodor commits to this fallacy is that he does not provide an independent justification for the notion of asymmetrical dependency. If he had done that, his asymmetrical dependency would have been the solution for the problem of misrepresentation.

Dretske laid out the fundamentals of his influential causal theory of content in his *Knowledge and Flow of Information*. His theory differs from Fodor's at least at four points:

- It does not assume a language of thought
- Causal effects of the external world on mental representations are interpreted as diachronic, i.e. the learning history of an organism is important for the content of a given mental representation
- It exploits Grice's distinction between natural and non-natural meaning
- It makes use of Shannon and Weaver's mathematical theory of communication, and thereby the notion of information.

Dretske uses the Gricean notion of natural meaning as the basis of his information theoretic semantics. The idea of natural meaning arises from natural signs. Natural signs mean something without any assistance from human beings. The existence of an oasis means that there is water around, the direction of a shadow means that the sun is in the other direction and so on. Unlike natural signs, non-natural signs are formed through some conventions among human beings, and they form the basis of non-natural meaning.

The relation between natural signs and their meanings is one of indication. Indication relations of this sort, as Grice points out, imply the factivity principle. That is to say, if an occurrence means (in a natural sense) that P, then it is the case that P. Dretske uses this property of natural signs for the function of a perceptual representation. Its main function is to indicate, or to carry natural information about, what is represented by the perceptual representation. In other words, if the representation is present, then the represented should also be present with a probability of one, and this is what binds behaviors to the external environment. Mental representations, which necessarily obey the factivity principle, are used in behavioral processes that are goal-oriented (or rather they have some specific functions). Representations contribute to these behavioral processes in virtue of their information carrying properties. Functions of behavioral processes and representations are constructed through a trial - error based learning history of the organism. Recognizing the importance of learning history makes Dretske's theory a diachronic theory as opposed

to synchronic theories such as Fodor's. Fodor's criticism of Dretske's learning period move is well known, especially in his review of correlational theories of mental content [Fodor 1984].

As in the case of Fodor's theory, misrepresentation is a problem for Dretske's theory. If the probability of the occurrence of the represented is 1 given the representation, how is it possible to have a misrepresentation that by definition implies the non-existence of the represented given the occurrence of the representation? Dretske in his earlier writings tried to solve this problem by a strict distinction between the learning period of an organism and the rest of its history (I will call these the learning phase and the retrieval phase respectively). An organism's trial and error based learning history fixes the content of a representation. After the learning period, he claimed, it is possible for a representation to acquire a content that deviates from its original function. Misrepresentation thus is possible only in the retrieval phase of an organism. After receiving immense criticisms regarding having no principled way of distinguishing the learning period from the mature period, Dretske took a teleosemantical turn in 1986. Instead of separating the learning period from the mature period of an organism, he used another unprincipled distinction: normal instances versus abnormal instances. The former ones, which are consistent with survival needs of an organism, fix the content of a representation, and the latter are allocated for misrepresentation. I examine Dretske's solutions in detail in the second chapter.

In short, none of the available causal theories provides a satisfactory theory of mental content. Despite the fact that they account for the referential dimension in accurate cases of representation, they have no way of explaining misrepresentation cases. The solutions offered by causal theorists for this problem turn out to be either based on unprincipled distinctions or a very clear case of question begging.

2. Conceptual Role Theories

Conceptual role theories are motivated by the inferential power of mental representations. Mental states influence our behaviors and other mental states via their inferential power. If I have a mental state that implies the existence of a cat, I will have a mental

state that has the content of a furry animal without any external assistance. For conceptual theories, mental representations are determined by conceptual roles like these, and these are all there is to the content of mental representations. Gilbert Harman (1982) and Brian Loar (1981) proposed theories in that line.

The main problem with such theories is the problem of communication or shared meaning [Fodor & Lepore 1992]. If the only factor that determines the content of my dog representation is its epistemic liaisons that I have at my mental disposal, then it is very likely that my dog representation is very different than yours. It is almost impossible for two people to share all the epistemic liaisons attached to one representation. 'They are dangerous' is definitely one of the epistemic liaisons of my dog representation whereas it does not have any connection with the epistemic liaisons of my ex-girl friend's dog representation. Put it differently, conceptual role theories imply a strict holism among mental representations, and holism is very vulnerable to the pure relativism objection.

The second main problem that conceptual role theories suffer from is about the truth conditions of a given representation. As mentioned above, the power of mental representations is a result of two dimensions: the inferential and the referential. Conceptual role theories take care of the inferential dimension by referring to the conceptual role whereas they are completely silent regarding the referential dimension since mental representations are cut off from the external world in their framework. When someone says that it is water by pointing to a glass of transparent liquor, he will be wrong. The utterance will be wrong because the referent of the mental representation that causes the utterance is not H_2O. Conceptual role theories are not able to give a satisfactory account of even such simple cases since the connection between mental representations and their referents in the external world is cut off.

3. Two Factor Theories

A natural reaction to the problem of mental content is that since mental representations derive their power from two main dimensions, then a satisfactory theory of mental content should combine these two dimensions in its framework. In other words, a mental

representation must have a causal connection to the external states of affairs as well as a conceptual role within the network of other mental states. Such a route has been taken by philosophers like Ned Block and Hartry Field.

These theories propose two independent components for explaining contentfulness of mental representations: referential and inferential. The former is supposed to explain the connection between mental representations and the external world. The latter is supposed to explain the inferential power of mental representations. Two-factor theories are attractive because the problems that one-factor theories are faced with could be solved by referring to the other independent component of two-factor theories. For example, the problem of misrepresentation, which causal theories severely suffer from, is dealt with in the inferential component. In a similar vein, the problem of shared meanings is taken care of within the causal component. In a sense, two-factor theories aim for the best of all possible worlds where they have an independent tool for each problem.

As Fodor & Lepore [1992] pointed out, the crucial problem for two-factor theories is the connection between the two independent components. On the one hand, these two have to be independent in two-factor theories in order to be able to solve the problems that other theories have. On the other hand, they are attached to mental representations that are complete entities. Something must make it possible for these two independent components to stick together in a given mental representation. Unfortunately, two-factor theories do not have a satisfactory answer for this.

4. An Interim Conclusion

A satisfactory theory of mental content is required for the needs of several different disciplines such as psychology, philosophy, linguistics and cognitive science. However, none of the theories that have been proposed is unproblematic and all of them are far from being satisfactory. Moreover, any satisfactory theory must be able to accommodate both the inferential and the referential dimensions from which mental representations acquire their power.

In light of these lessons, I will provide five preliminaries that are necessary, but far from being sufficient, for a satisfactory theory of mental content in the following section.

5. Preliminaries for A Theory of Mental Content

As mentioned above, mental content has two important dimensions that need to be explained. Emphasis on the external dimension leads to the problem of misrepresentation, and emphasis on the internal one brings the problem of shared meanings and the problem of being cut off from truth conditions. The problems that are caused by the latter emphasis are more severe than the problem caused by the former one. Without being able to satisfy a ground for shared meaning, communication and any linguistic representation would be impossible. On the other hand, the problem of misrepresentation, despite its importance, does not cause any devastating effect like these. Moreover, it may be possible to accommodate the possibility of misrepresentation within a causal framework. In fact, I explore this option in this dissertation and construct a theory based on a causal/informational approach that does exactly that.

Communication is essential for developing any kind of mental ability. This fact is true not only for humans but also for animals. Communication here refers to a wider range that includes animal communication as well [7]. The cases of feral children [Mason 1972] provide a justification for this claim. Such children who grew up in an environment where no social interaction is available have been reported for several hundred years. When these children are found in the wilderness and brought back to the 'civilization', where communication is fundamental, they could not adjust to the new environment. They could not learn to speak beyond a couple of simple words, and they had big difficulties in learning some basic activities like walking, eating and getting used to being dressed up. These cases show us that social interaction and communication are essential for having mental abilities that

[7] Machine communication is deliberately left out.

are utilized by mental representations.[8]Hence, a theory of mental content that provides an explanation for shared meaning and communication is preferable to those which do not.

The ability to account for truth conditions is another essential feature of a satisfactory theory of mental content. Besides the philosophical importance of such an ability, which is well presented in Fodor & Lepore [1992], there is empirical evidence for the importance of such a connection for learning and the development of mental representation. Learning is not possible without feedback and control mechanisms that human beings are endowed with. Describing both of these mechanisms requires us to identify the truth conditions of a mental state. Additionally, if the human mind is cut off from the external world, i.e. no truth conditions, then the mind will have to live in its darkness. New developments in cognitive science also show us the importance of the external world for human mind. The growing literature on the embodied mind approach is a result of the need of putting the external world back into the picture. Theories of mind after oscillating between two extremes, fully internal introspectionism and fully external behaviorism, have been trying to find a balance by adding first mental representation (cognitive psychology) and then the external environment (embodied approach) into the picture [Gardner 1984; Clark 2000].

The importance of the ability to account for shared meaning and truth conditions takes us to our first preliminary.

Preliminary 1: the referential power of mental representations is a better starting point than the inferential power for a satisfactory theory of mental content since it provides room for communication.

The referential dimension of mental states expresses the connection between mental entities and the external world. The relation between these two is one of representation. Two properties of the representation relation have proven to be essential: singularity and

[8]In this passage, I do not mean that feral children have no mental abilities at all. Apparently, they have some form of mental abilities. This is exactly why they survived in the wilderness. However, the mental abilities that they have do not qualify as the mental content that is of interest to us. The only point her is that for mental content with all of its essential properties, social interaction and communication are necessary.

asymmetry [Fodor 1992]. My mental state representing my computer represents the very computer that I am using right now; it does not represent any other computer until the singularistic representation becomes a concept of computer. Hence, the relation between my mental representation and the object in front of me is singular. The relation is asymmetrical, too. The mental entity that I have for my computer represents the computer that I am using right now, but my computer does not represent the mental representation that I have at my mental disposal. One can think of the representation relation as a two-place relation: X_1 represents Y_1 (RX_1Y_1). The relation is not symmetrical, to wit, the fact that X_1 represents Y_1 does not necessarily imply that Y_1 represents X_1.

Any theory that aims to explain the nature of the relation between mental representations and the external world must be consistent with these two basic properties.

The most intuitive approach to the nature of the relation between mental representations and the external world is one of resemblance (or similarity). My mental representations resemble the external states of affairs, and such a resemblance is the basis for explaining the possibility of shared meaning across individuals and for accommodating truth conditions for my representations. This intuitive approach, however, is not consistent with two basic properties mentioned above. First of all, resemblance (or similarity) is a symmetrical relation. X resembles (or is similar to) Y implies that Y resembles (or is similar to) X. Moreover, it is not possible to accommodate singularity in such an approach either. If my representation of my computer is formed via resemblance to the computer in front of me, then my representation will resemble other computers, and as a result my mental representation will be about other computers, too. My current mental representation will loose its singularity. Hence, two basic properties of the representation relation cannot be accounted in resemblance-based theories. [Fodor 1992; Cummins 1991]

A causal approach, as opposed to non-causal approaches such as resemblance-based theories, is consistent with both asymmetry and singularity of mental representations. If an entity, which could be just a property or a whole object[9], in the external world causes

[9]I prefer not to specify the nature of entities that can cause mental representations. It does not matter if they are objects or just properties at this point.

my mental representation of that specific entity, then that entity will not represent my mental representation, since my mental representation is not the causal origin of the external entity. Thus, the relation is asymmetrical. For singularity, my mental representation of my computer, say R, is caused by the computer in front of me and it is not caused by any other computer, so R represents only my computer and no other computers.

Besides the ability to account for asymmetry and singularity, causal approaches are consistent with the empirical findings regarding sensory mechanisms, especially in vision research. For example, if a certain part of the visual cortex of a frog gets activated when the frog sees a fly, it is inferred that the part in question is a motion detector and the neuronal activation in that part is caused by the existence of a fly in the visual field of a frog [Cummins 1996, p.9]. Hence, a causal approach for the nature of the representation relation is more in line with our scientific understanding.

These remarks show us that the nature of the relation between mental representations and the external world is better explained by a causal approach than non-causal approaches. Hence, the second preliminary

Preliminary 2: a causal approach is a good candidate for the nature of the referential power of mental representation.

Causal approaches imply the following: an external state of affairs, S, causes a mental entity, R, and thereby, R represents S. Such a representation relation makes it possible for human beings to deal with states of affairs that fall under the same type with S. In a sense, we learn something about S by having R. This could be a further step for understanding the nature of representation relation. After all, it is common to learn about causes from their effects.

Following such reasoning, Stampe offers an epistemic account of the representation relation. He says,

> 'An object will represent or misrepresent the situation ... only if it is such as to enable one to come know the situation, i.e. what the situation is, should it be a faithful representation.' [Stampe 1977]

Although this claim has some truth to it, it has trouble with two basic properties of representation relation mentioned above: asymmetry and singularity. The epistemic account makes the representation relation symmetrical. For example, the barometer represents the weather, but the weather does not represent the barometer. However, in the epistemic account, both should represent each other since it is possible to learn about the weather from the barometer as well as to learn about the barometer from the weather. If it is raining outside, the barometer is probably low.

Singularity is also a problem for the epistemic account. Stampe uses a portrait of Chairman Mao as an example. If the portrait is faithful, which is a necessary condition for a representation relation in Stampe's account, we can learn a lot about Chairman Mao's physical appearance, even about his personality by analyzing his facial features. The possibility of learning these is another piece of evidence that the portrait represents Chairman Mao. However, if Mao had a doppelganger, then we can also learn about his doppelganger from the portrait. Thus, the epistemic account implies, the portrait will be representing both Chairman Mao and his doppelganger. Such a result violates the singular nature of the representation relation. [Stampe 1977]

The inability to account for singularity and symmetry rules out the epistemic account as a successful explanation for the nature of representation. However, this does not imply that it is not possible to learn from mental representations about the states of affairs that cause them. On the contrary, mental representations are the main vehicles by which we gather information about the external world. The problem is that the learning feature of representations should be explained via another relation that is consistent with singularity and asymmetry. The best way of finding such a relation is to analyze natural causation instances. It is a natural law that the existence of fire causes the existence of smoke. As a result of such a law, the existence of smoke tells us something about fire. In a sense, smoke indicates fire in a natural way. Using Gricean terminology, smoke is a natural sign for fire. The causal law between fire and smoke, fire causes smoke, brings about the indication relation between smoke and fire: smoke indicates fire. However, fire does not indicate smoke. Thus, the relation is asymmetrical. It is also singular, because the smoke coming

out of my cigarette indicates the specific fire that is burning the particular cigarette that I am smoking now, and not any other fire. Since the indication relation is based on a causal law, it is possible to learn about the cause from the sign that bears an indication relation. The indication relation is consistent with singularity and asymmetry of representations, and moreover it encompasses the possibility of learning from effects about their causes.

These remarks make the indication relation a good candidate for understanding the nature of the representation relation between mental representations and the external world. It is important to note, however, that the indication relation does not bring a problem free landscape for theories of mental content. It has one main problem. Indication works well in the case of natural laws, but a theory of mental content must explain non-natural cases of the representation relation too. It is almost impossible to explain convention based representation instances (for example linguistic representation) by the indication relation. The reason for this is very simple: whereas misindication is not possible, misrepresentation is an essential feature of non-natural representation instances. To put it slightly differently, it is a challenge to construct meaning out of the indication relation. Nevertheless, this problem should not make us lose our optimism regarding indication. Following Grice's and Dretske's footprints, I think that it is possible to specify the connection between natural signs (indication) and non-natural ones (representation). Hence, the third preliminary

Preliminary 3: the indication relation in the case of natural laws is consistent with two fundamental properties of the nature of representation relation. The challenge is to generalize such a relationship to non-natural cases of representation such as mental representation and linguistic representation.

The preliminaries that are covered so far are about the referential dimension of mental representation. As mentioned above, mental representations have their power over behavioral and mental states from the inferential dimension as well. A satisfactory theory of mental content, for which the referential dimension is a better starting point, should provide a satisfactory ground for the inferential (transformational) power of mental representations too.

The common way of dealing with the inferential power of mental representations within causal approaches is to exploit syntactical properties of representations. There is a strong correspondence between form and content of representations in computational systems. The content of a representation is fixed via the causal connection with the external world. Contents of different representations portray a specific relation, and this very relation corresponds to the relation between the syntactical properties of the representations in question. The following quote from Fodor explains this idea.

> You connect the causal properties of a symbol [mental representations are symbols for Fodor] with its semantic properties via its syntax. The syntax of a symbol is one of its second order physical properties. To a first approximation, we can think of its syntactic structure as an abstract feature of its (geometric or acoustic) shape. [Fodor 1992, p.22]

The parallel structure between syntactical (computational) and semantical (cognitive) properties of mental states is dubbed as *the Tower Bridge Structure* by Cummins [1996], and is criticized by him because of the possibility of misalignment between two parallel levels.

The tight correlation between cognitively relevant content level and computationally relevant form level is threatened by two familiar sorts of cases. First, Frege cases where the same content could be expressed by different forms, morning and evening stars both refer to Venus. Second, by twin-earth cases where the same form is attached to different contents, the concept of water is attached to H_2O on the Earth and to XYZ on Twin Earth. Fodor himself accepts that these cases violate the tight correspondence between semantical and syntactical levels, and he uses these cases for criticizing two-factor theories. On the other hand, he also needs in his theory the very correlation that is under his attack. His solution for these cases is hardly a solution, he says that these cases do exist, but they are rare and controlled by specific mechanisms of the human mind.

Cummins claims that misalignment cases, after all, are not that rare and that the arbitrariness of primitive representations in causal theories is the main problem. As long

as these are arbitrary, they could get attached to any content and such an attachment will lead to misalignment in quite a number of cases.

The misalignment problem decreases the plausibility of a tight correlation between content and form for explaining the inferential power of mental representations. Causal approaches need to find another entity that can glue the internal and the external dimensions together. The notion of information, I claim, could be the savior at this point. In information-theoretic causal approaches, representations carry information about the world. Such a claim is consistent with our common sense intuitions and our empirical understanding of sensory mechanisms. Cummins explains the plausibility of the claim better than I could, so I quote him.

> If you have a cell that fires when and only when it encounters a visual edge, then those firings carry the information that there is a visual edge present. And since it is evidently useful to have information, it seems plausible to suppose that representing is just indicating. [Cummins 1996, p.64]

Such a use of the notion of information is definitely plausible and could be used in a promising theory of mental content because of its close connection to the indication relation. However, this is not the only feature that makes the notion of information attractive. The notion of information can also explain the inferential (transformational) power of mental representations. The information carrying property of representations is in parallel with both the referential and inferential dimensions of mental representations.

There are three more reasons for why the notion of information is promising for a successful theory of mental content: i) it provides a level of abstraction different than causal interactions; ii) it carries a primitive form of intentionality; iii) the success of the mathematical formalism of Shannon's theory. For the first point, there are indefinitely many different sensory inputs that give rise to the same cognitive outcome. So, causation is not enough to generalize over all these different sensory inputs. We need a higher level

of abstraction, and the notion of information seems to be a good candidate for the job. Dretske explains this a lot better than I could.

> For we then discover that there are an indefinitely large number of different sensory inputs, having no identifiable physical (non-relational) property in common, that all have the same cognitive outcome. The only way we can capture the relevant causal regularities is by retreating to a more abstract characterization of the cause, a characterization in terms of its relational (informational) properties. [Dretske 1983, p.76]

Secondly, information carries a primitive intentionality. One could have the information that Susan is my sister without having the information that she is Harry's husband (just like beliefs). On the other hand, causation does not cut in this way. Since Susan is the cause of my mental representation of hers, and since she has both properties, there is no way of separating one property of hers from her other properties.[10]

Thirdly, the mathematical theory of communication is well developed and it exploits the statistical relations between a source and a receiver. After all, the world is the source and our perceptual mechanisms are the receiver, and we exploit the statistical properties of the world. So, why not utilize the technical tools that are developed for that purpose.

These remarks take us to the fourth preliminary.

Preliminary 4: The notion of information, if included in a theory of mental content, could combine the inferential and the referential dimensions of mental representations. Moreover, it provides a primitive form of intentionality

[10]It should be noted that I am using the notion of causation in a general way. There are different theories of causation such as singular event causation and property causation. It is true that singular event causation does not provide the primitive intentionality that the notion of information provides. However, this is not at all clear for property causation theories. It is an important question whether or not property causation can provide a primitive intentionality. Unfortunately, I cannot delve into this question in this project. To answer this question would obviously require another dissertation. I thank Prof. Frederick Schmitt for pointing that out

As we have seen above, to account for misrepresentation cases is required as well as to account for accurate representation cases. As clearly expressed by Fodor, conditions that determine what a mental entity represents must be dissociated from the truth conditions for that mental entity [Fodor 1992, p.42]. In other words, conditions that determine the content of a mental representation must be close enough to its source in order to account for the referential dimension. However, those conditions should not be too close to the source either, since the dissociation between content assignment conditions and truth conditions is required for explaining misrepresentation cases. Cummins & Poirier [2004] contrast representation and indication relations. They show that indication relations, despite being a natural starting point for describing representation relations, have different characteristics than representation relations. The main difference lies in the fact that indication is source dependent whereas representation is not. Therefore, any theory of mental representation must provide a way of source independency. In their own words,

> Indicators are source dependent in a way that representations are not. The cells studied by Hubel and Weisel all generate the same signal when they detect a target. You cannot tell, by looking at the signal itself (the spike train), what has been detected. You have to know which cells generated the signal. This follows from the arbitrariness of indicator signals, and is therefore a general feature of indication: the meaning is all in who shouts, not in what is shouted. In sum, then, indication is transitive, representation is not. It follows from the transitivity of indication that indicator signals are arbitrary and source dependent in a way in which representations are not ... [Cummins & Poirier 2004]

Hence, the last preliminary ...

Preliminary 5: In order to account for misrepresentation cases, or in other words in order to satisfy the source independency of representation relation, the truth conditions and the content assignment conditions of mental representations must be dissociated.

Given these five preliminaries, Dretske's 1981 framework, which is briefly described in Section 1, seems to be the best alternative available. The Dretskean framework satisfies four of five preliminaries. It starts out with nomic relations between a mental entity as a sign and the external state of affairs as its signified, and thus satisfies the referential and causal requirements of Preliminaries 1 and 2. Moreover, it uses the Gricean distinction between natural vs. non-natural meaning. As mentioned above, natural meaning instances in Grice's theory fall under the category of indication relations. Hence, the Dretskean framework satisfies the the third preliminary as well. As for the fourth preliminary, which is about the importance of utilizing the notion of information, it is clear that Dretske aims to use the notion of information and some concepts of the mathematical theory of communication such as entropy in his theory of mental content. However, Dretske's theory fails to fulfill the requirements of the fifth preliminary. He uses the notion of informational content for defining mental content, and in his definition for informational content he assigns unity to the probability of the relevant external state of affairs given a sign (a mental representation). This move makes the representation relation too close to its source, hence it is impossible to dissociate the truth conditions from the content assignment conditions. Here is how he defines the notion of informational content.

> **Informational Content:** A signal r carries the information that s is F = the conditional probability of s's being F, given r (and k), is 1 (but, given k alone, less than 1) [k refers to background knowledge] [Dretske 1981, p.65]

Thus, his theory does not fulfill the requirements of the fifth preliminary, to wit, the problem of misrepresentation afflicts his theory. Despite the fact that this claim is commonly accepted and well-justified, to do justice to Dretske's 1981 theory requires analyzing the solution that Dretske offers for the problem of misrepresentation. This is the task that I undertake in the next chapter.

CHAPTER 2

Misrepresentation: Dretske's Solution is on Trial

In his 1986 paper, *Misrepresentation*, Dretske suggests a solution for the problem of misrepresentation within his information theoretic semantics. This chapter's overall aim is to analyze Dretske's solution, and to show why the suggested solution does not work.

Loewer [1987], in his article *From Information to Intentionality*, argues against Dretske's proposed solution for the problem of representation. I agree with Loewer's final judgment, but for different reasons. Loewer's justification for his claim is based on a misunderstanding of Dretske's approach. In the second section of this paper, I critically explore Loewer's ideas, and attempt to argue that he misunderstands.

1. Dretske's Solution

1.1. Natural meaning and Functional Meaning. Dretske uses Gricean notion of natural meaning as the basis of his information theoretic semantics. The idea of natural meaning arises from natural signs. Natural signs mean something without any assistance from human beings. The existence of an oasis means that there is water around, the direction of a shadow means that the sun is in the other direction and so on. The main motivation for introducing natural meaning is to differentiate naturally occurring signs, which do not require any assistance from us, from non-natural signs that are formed through some conventions among human beings. Any natural language or any code of communication is a good example of non-natural signs and non-natural meaning.

Dretske, following in Grice's footsteps, claims that natural meaning implies its truth. To put it differently, if an occurrence means (in a natural sense; henceforth mean$_n$) that P, then it is the case that P. This property of natural meaning is called the factivity principle in Gricean terminology. A very commonly used example of the principle is the indicative

relationship between 'red spots on the face' and 'having measles'. There is a lawful relation between red spots on the face and having measles. This lawful relationship is what constitutes the natural meaning of red spots. However, in some instances red spots might be caused by other factors. In such situations, according to Dretske and Grice as well, the symptom of having red spots loses its natural meaning. In other words, the red spots on Tommy's face mean$_n$ that Tommy has measles only if Tommy really has measles.

This property of natural meaning leads Dretske to a particularistic account of meaning. Natural meaning is accepted only in particular instances, to wit, there is no type-associated meaning of the symptom of having red spots on the face. We cannot talk about the meaning of having red spots on the face generally in a natural sense, according to Dretske. Whenever we talk about this symptom and its relation to having measles we should be talking about particular instances in which the person really has measles. In his own words,

> In speaking of ... natural meaning I should always be understood as referring to *particular events, states or conditions*: this truck, those clouds, and that smoke. (Emphasis is original) [Dretske 1994, p.159]

Although Dretske accepts the possibility of defining a type-associated meaning in a natural sense, he sees no use of such a possibility in solving the problem of misrepresentation. Before proceeding with the rest of my discussion, it is worth pausing and contemplating on the implications of not having a type-associated natural meaning. Imagine that you are looking at two people, Tommy and Alice, and both have red spots on their faces, but only Tommy has measles. The following three statements about this situation are compatible in Dretske's account.

(1) The red spots on Alice's face do not mean$_n$ that she has measles.
(2) The red spots on Tommy's face mean$_n$ that he has measles.
(3) Although the red spots on Tommy's face mean measles, having red spots as such does not mean measles.

Once the stage for natural signs (meaning) is set up in this way, then it is not possible to embrace the possibility of misrepresentation by them, because 'they either do their job or

they do not do it at all.' On the other hand, what is needed for misrepresentation is a kind of meaning that admits the possibility of not fulfilling its task.

At this point, one should ask the status of the functions of human made devices and their relation to natural meaning. What does, borrowing Dretske's example, a ring on the doorbell mean? We are all inclined to say that it means that 'there is someone at the door'. This inclination of ours is a result of the design of the doorbell, thus, it is *supposed to mean*$_n$ that there is someone at the door. Although Dretske accepts this supposed natural meaning, he denies that the ring on the doorbell *actually means*$_n$ that 'there is someone at the door'. This denial, once again, is a result of his particularistic account. In order to be able to talk about natural meaning, we must use particular instances, and since in some instances the ring may be the result of a short circuit, we cannot say 'a ring on the doorbell as such means$_n$ that there is someone at the door' in such instances.

Ironically enough, although his presuppositions about natural meaning necessitate this denial, he still wants to be able to talk about the meaning of a ring on the doorbell as such. He says,

> Granted, one may say, the doorbell's ringing cannot mean$_n$ that someone is at the door; still, *in some related sense of meaning*, it means this whether or not anyone is there. If this is not natural meaning (meaning$_n$), it is a close cousin. (Emphasis is added) [Dretske 1994, p.160]

He dubs this close cousin of natural meaning as functional meaning. Whenever there is an identifiable function attached to a device and/or an organism, then one can talk about functional meaning of the given state of a device or an organism. For example, the position of the needle of a fuel gauge tells us the amount of gas in the tank. If it points to the far left, it means that the tank is empty. When the mechanism of a fuel gauge is working properly (i.e. when the tank is really empty), we are eligible to use natural meaning, according to Dretske. Whenever it is not, then the natural meaning is out of the question because of the aforementioned factivity principle. However, Dretske claims, there is *a sense* in which the needle's position to the far left means that tank is empty whether or not the tank is really

empty because the fuel gauge is designed in this way. And this *sense* is 'functional meaning'. To put it differently, when the needle is at the far left, it is always *supposed to mean*$_n$ that the tank is empty. This supposed mean$_n$ is what it means$_f$ (functional meaning). Let me quote Dretske for the formal definition of functional meaning.

> **Mf:** d's being G means$_f$ that w is F = d's function is to indicate the condition of w, and the way it performs this function is, in part, by indicating that w is F by its (d's) being G. (Key. d: the device, G: a state of d, w: the world, F: a state of the world) [Dretske 1994, p.161]

By introducing functional meaning, Dretske creates a type-associated meaning that the possibility of which he denied for natural meaning. A theory of meaning without a type-associated meaning beyond and above particular instances is counter intuitive. We all, including expert physicians, want to say that having red spots on the face as such generally means that the person has measles. If the reader remembers, Dretske's particularistic account implied the falsity of such statements at the beginning. However, since he needs to have something to hold on to, he introduces 'functional meaning'. Functional meaning remedies the lack of type-associated meaning in the case of functions. However, it cannot do so in the cases of medical symptoms, animal communicative behavior and etc. Although this is one of the main shortcomings of Dretske's semantics, I shall discuss it elsewhere given the subject matter of this chapter, which is the problem of misrepresentation.

The other motivation that led Dretske to introduce a new sense of meaning is to avoid the trap of the factivity principle since it does not make room for misrepresentation cases because of its very definition. As mentioned above, natural meaning implies the factivity principle, and therefore, cannot accommodate the possibility of misrepresentation. Dretske hopes to fix this problem by introducing the notion of functional meaning.

1.2. Misrepresentation. The first step in embracing misrepresentation in a theory of semantics is to make room for false content, which is a necessary condition for misrepresentation, but not a sufficient one. The ability of having false content does not guarantee

having false beliefs. The latter requires more. Only when false content has all the properties of propositional attitudes can one speak of having false beliefs. However, without false content it is not possible to talk about false beliefs at all. Thus, one must make room for false content first in his theory of semantics and Dretske does so by using the notion of functional meaning.

Functions can come in different ways. We may assign functions to the states of devices and organisms. For example, we assign the function of showing the amount of fuel in the tank to fuel gauges or we assign the presence of marijuana to the barks of some trained dogs, etc. Unlike these assigned functions, there are functions of natural states of organisms. For example, one of the functions of the eye blinking state of my body is to protect my eyes; this function is not assigned by someone else. In order to have a naturalized theory of semantics, one should use natural functions, not assigned functions. The main danger of using the latter is to cause circularity because they are assigned by organisms with intentionality, namely by human beings. If we were to use them then we would be in a situation where the false content is explained by referring to some other agent's intentional capacity. Hence, one must use natural functions not assigned ones. The obvious place to look for natural functions is simple natural organisms and the functions of their internal states which are evolved as a result of their natural needs.

Because of these constraints, Dretske chooses a marine bacterium as his primary example, and analyzes its ability to have false content. Before proceeding with the details of the marine bacterium in question, I should mention that since Dretske wrote his article we have learned a lot more about the sensory mechanisms of the marine bacterium in question. Apparently, its sensory mechanisms are much more complicated. However, introducing these new complexities would shadow Dretske's original argument. Therefore, I choose to stick with Dretske's original presentation. The details of the complexities of the marine bacterium could easily be found on the web page of one of its discoverers, Richard B. Frankel (www.calpoly.edu/ rfrankel/magbac101.html).

This marine bacterium moves towards the magnetic north (or south depending on which hemisphere it is in). The survival value of this continuous move is to avoid surface water

where the oxygen level is dangerous for the life of the bacterium. It moves towards geomagnetic poles via a mechanism called magneto taxis. The mechanism detects the geomagnetic field that points to the North Pole (or the South Pole). Upon the detection, the bacterium moves in the direction of the magnetic field. Hence, it is able to avoid oxygen-rich surface water, proceed to an environment with comparatively less oxygen and stay alive. Now, the main question is what is the functional meaning of the magneto taxis mechanism. The natural answer to this question is to point to the direction of comparatively oxygen free water (let's symbolize this function with F_1). It is possible to get false content under this assumption. The bacterium could be directed to a deadly environment by putting a bar magnet oriented in the opposite direction of the geomagnetic field into its environment. This scenario is very promising for having a false content, because we have a content that does not correspond to the external state of affairs, because the bacterium moves towards relatively oxygen-free water contradicting the function of its magneto taxis mechanism. Therefore, our bacterium has false content about its surroundings.

This story sounds very plausible so far, but Dretske questions the assumption about the function of the magneto taxis mechanism. This mechanism works properly even in the story of lured bacterium, so there is no malfunctioning. Maybe its function is different; maybe its function is just to point in the direction of the magnetic field (let's symbolize it with F_2) rather than pointing to a relatively oxygen-free environment. If this is the case, then there is no way of claiming that we have false content in the lured bacterium example, because the mechanism properly fulfills its function, i.e. to point in the direction of a magnetic field. Dretske claims that there is no good reason for preferring F_1 to F_2. If we change the function that we assigned to the magneto taxis mechanism, then there is no false content in the bacterium's 'internal world', the disaster of moving to a deadly environment is to be blamed on the external world rather than the bacterium's false content. Therefore, he infers, it is not possible to get a false content with our simple bacterium. Let me summarize Dretske's argument in a more formal way.

(1) There two possible functional meanings that could be identified for the magneto taxis mechanism: F_1-to point to an environment where the oxygen level is not at odds with the survival needs of the bacterium or F_2-to point to the magnetic field, not necessarily the geomagnetic field.
(2) If the function is F_1, then the bacterium has false content, i.e. 'representing' the external world falsely, when it is lured to a deadly environment with a bar magnet.
(3) If the function is F_2, then there is no false content let alone misrepresentation even in the luring scenario.
(4) We have no good reason for preferring F_1 to F_2.

Hence,

(5) There is no place for false content in the functional mechanism of the bacterium.

Although his argument seems to work at first glance, the fourth premise is questionable. This premise arises because of Dretske's claim of the indeterminacy of function of the magneto taxis mechanism. It could be either of the two functions mentioned above.

Dretske chose the magneto taxis mechanism as the paradigmatic example, because it is the result of a natural need of an organism. The natural need of the bacterium in question is to move towards deep water where the oxygen level is not life threatening. Had it not had this need then the evolution would not have given the gift of the magneto taxis mechanism. Thus, it seems to me that we have a very good reason for choosing F_1 over F_2 as the function of magneto taxis mechanism. If this is true, then we should reject the fourth premise of Dretske's argument. In addition to that, we should accept that the bacterium's internal state has a false content in the luring experiment, because this is exactly what Dretske claims in the second premise of his argument. We identified the function of the magneto taxis mechanism as being to point to relatively oxygen free water because of its survival needs, and in the luring scenario, the mechanism does not fulfill its function. Although, I think that this is sufficient for refuting Dretske's argument, it is still useful to continue analyzing his reasoning after the bacterium example.

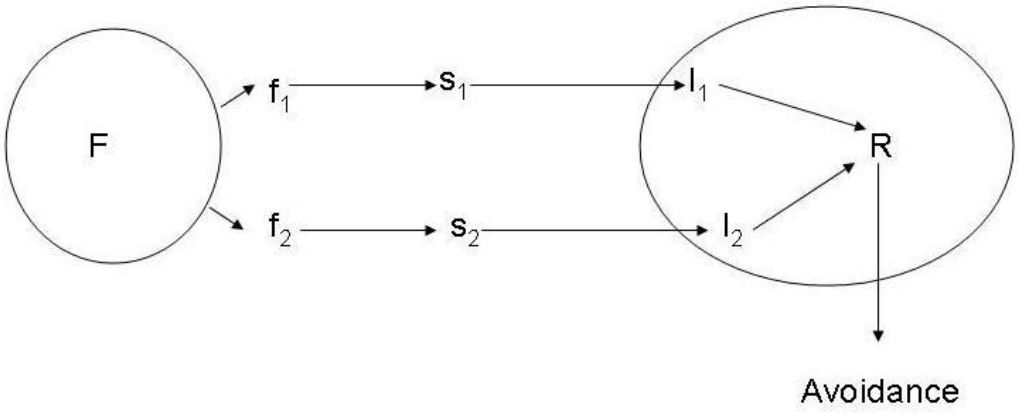

FIGURE 1. Dretske's Hypothetical Organism

As a further claim to his argument, Dretske says that it is not surprising not to have room for false content in the case of the bacterium because of its simple structure. A certain threshold of complexity, he says, is required for false content and thereby for misrepresentation. In his own words,

> For the purpose of clarifying issues, I have confined the discussion to simple organisms with primitive representational capacities. It is not surprising, then, to find no clear and unambiguous capacity for misrepresentation at this level. For this power requires a certain threshold of complexity in the information processing capabilities of a system.[Dretske 1994, p.165]

In order to prove his point, he stages a hypothetical organism, O, and analyzes two different versions of it for the ability of having false content.

The first version of Dretske's organism O is able to detect a toxic substance, F, by two different sense modalities or by exploiting two different signals within a single sensory modality. The properties of F, f_1 and f_2, give rise to two stimuli, s_1 and s_2, that are perceivable by the sensory mechanism of the organism. As a result of perceiving these two proximal stimuli, the organism gets two different internal states, i_1 and i_2, and both of these internal states lead to another state R that means nothing but 'avoid'. Hence, the organism runs away from F and by doing so it stays alive. Figure 1 summarizes the situation.

Since Dretske wants to base the ability of having false content on functional meaning, it is crucial for him to identify the function of the mechanism. Intuitively, it seems that the function of this organism is to detect the existence of F. If this is correct, then the organism has false content when it is deceived with an ersatz F. Let's see this deception story in detail.

We deceive the organism with an ersatz F as we lured the bacterium with a bar magnet, and say the ersatz F gives rise to s_1. Given the basics of the story, s_1 triggers a chain of events i_1, R and avoidance. What is the function of the mechanism in this case? Dretske says nothing short of detecting F would work as an answer, because the occurrence of R does not mean s_1, it could have been s_2 since we have two different properties, s_1 and s_2, that can trigger the chain of events that end up with R, and avoidance is a natural result of R. So, he says, the functional meaning of this new organism, as compatible with our intuition, should be detecting F.

Contrary to our intuitions, Dretske adds, the idea of the indeterminacy of function is in play in this example, too. The function of the organism could very well be to detect s_1 or s_2 (a disjunctive meaning) rather than to detect F. Under this assumption, however, to embrace false content (and thereby misrepresentation) fails. The organism is able to detect s_1, which is a result of ersatz F not real F, and since its function is to detect s_1 or s_2, it fulfills its function properly, hence no false content.

After raising this point, he applies the same reasoning that he used for the bacterium example, and concludes that there is no good reason for preferring one function to another: detecting F and detecting s_1 or s_2. If this is correct, then there is no room for false content in this example either. His argument goes as follows.

(1) There are four possible functional meanings that could be assigned to the detection mechanism:

to detect s_1

to detect s_2

to detect F

to detect s_1 or s_2.

(2) The former two cannot be the case that since the detection mechanism can be triggered with either of s_1 and s_2.

(3) If the function is to detect F, then the organism has false content, i.e. 'representing' the external world mistakenly, when it is lured with an ersatz F.

(4) If the function is to detect s_1 or s_2, then there is no false content let alone misrepresentation in our luring by ersatz F thought experiment.

(5) We have no good reason for preferring 'to detect F' to 'to detect s_1 or s_2'.

Hence'

(6) There is no place for false content in our organism.

As in the bacterium example, the fifth premise is problematic. The organism in question developed the detection mechanism for its survival needs, and since the natural functions, as Dretske himself says, are based on natural needs we have good reasons for identifying the function of the mechanism as being to detect F, not s_1 or s_2. Despite this fact, Dretske insists on the notion of indeterminacy of function, claims that his complex organism is not able to have false content unless we make a change ...

The change that he proposes is related to the level of complexity of the organism. He expands the complexity level of the organism by adding an associative learning mechanism. The second version of his organism can learn to detect F with new stimuli as the organism exposed to F's presence. In other words, the organism improves its ability to detect F by learning its new properties. For example, at time t_1, only s_1 and s_2 can trigger a chain of events that end up with R and avoidance whereas after a learning period at time t_2 it is able to detect F by a new property say s_3. Figures 2 and 3 present an example of this learning ability.

Dretske also allows the associative learning mechanism to work with stimuli that have nothing to do with the properties of F. Any arbitrary stimulus could trigger the organism's avoidance behavior if that stimulus is repeatedly exposed to the organism at the presence

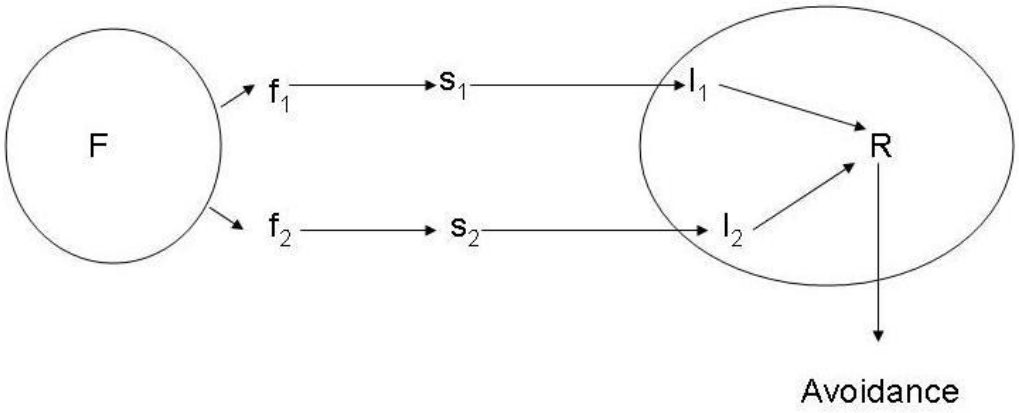

FIGURE 2. Learning - Time t_1

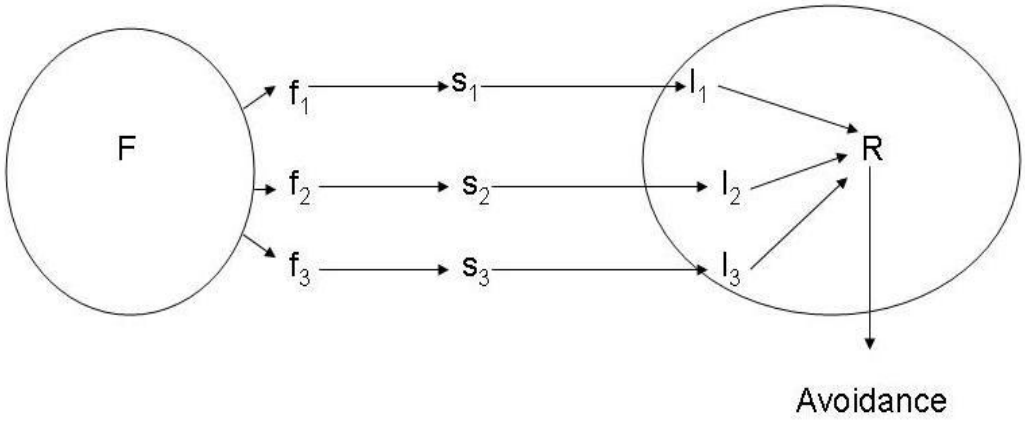

FIGURE 3. After a Learning Period - Time t_2

of F. In other words, the organism could be conditioned with any kind of stimuli, even with those that have no connection to F and its properties.

In this complex version of Dretske's hypothetical organism, the natural meaning of R, a state of the complex organism, changes over time. At time t_1, R means$_n$ s_1 or s_2; it means$_n$ s_1 or s_2 or s_3 at time t_2. Thus, the natural meaning of R depends on the individual's learning history. It varies over time as well as across individual organisms. There is no time-invariant natural meaning.

Despite the continuous change in natural meaning, Dretske says, R keeps indicating the presence of F. A newly learned property of F will be correlated with the presence of F, and any conditioned stimuli will indicate its presence, too, because this is the main idea behind conditioning. Even these conditioned stimuli will become the natural signs for F at a given time. So, if we want to assign time-invariant functions to cognitive mechanisms such as the organism O with associative learning capability, we need to accept that the function is to detect the presence of the source (the presence of F in our example). Otherwise, we will not be able to find any time-invariant function for any cognitive mechanism; thus, the function of the state R of the organism is to detect the presence of F, according to Dretske. This is how he eliminates the indeterminacy of function.

Given the above ideas, Dretske is ready to accept the ability of having false content for the revised version of the complex organism. Since the function of the state R is to detect F and when this very state is triggered by an ersatz F, then R will have false content about the state of the affairs of its surrounding.

The main problem with Dretske's reasoning is that he makes a huge jump from the claim that there is no time-invariant natural meaning to there must exist a time-invariant functional meaning. He sees no problem with the continuous change in natural meaning of a state of an organism. On the other hand, he wants to have something to hold on to over time for functions. His reasoning is not plausible unless he gives a justification for the difference between natural meaning and functional meaning. Unfortunately, he does not mention any such justification in his article.

As mentioned above, the main motivation of proposing a new type of meaning, functional meaning, is to have a type-associated meaning for Dretske since his notion of natural meaning is a strictly particularistic one. So, to seek for a time-invariant functional meaning is consistent with his original motivation.

1.3. An Interim Conclusion. In summary, Dretske denied the possibility of false content in both the primitive bacterium example and the first version of his hypothetical organism. His reason for this denial is the indeterminacy of the functions of the mechanisms

of these examples; magneto taxis and detection of F, respectively. For example, one can identify the function of the magneto taxis mechanism of marine bacteria as being to point in the direction of magnetic field, rather than the direction of geomagnetic field. And, as a result, the false content cannot be embraced even when the bacterium is lured by a bar magnet. Dretske contradicts with himself in this reasoning. As he himself states, choosing natural mechanisms based on natural needs as a starting point for embracing misrepresentation is a necessary condition for a naturalized and non-circular semantical approach. Since these natural mechanisms are developed because of their survival values for the organisms, it is an expected result to choose the functions that have survival values to other functions that do not. In the case of bacterium, it is reasonable to say that the function of the magneto taxis mechanism is to point to the geomagnetic field, and thereby deep water, instead of saying the function is to detect any kind of magnetic field. The survival value of a mechanism does matter, and that should be used to identify the function of a mechanism.

The only case (among the three examples he discusses in his article) for which Dretske accepts the ability of having false content is the organism with associative learning ability. He claims that in this case the problem of the indeterminacy of function can be eliminated, since the natural meaning changes continuously, and we need a time-invariant function for the states of cognitive mechanisms. As I said earlier, he mentions this need without any justification. What he claims comes down to the following. *We can get rid of the indeterminacy of function in this example because we need a determinate function.* I do not think that much can be said for such a claim.

2. Loewer's Response

In his article, *From Information to Intentionality*, Loewer [1987] discusses Dretske's account of misrepresentation suggested in his 1986 paper. In this article, Loewer revises Dretske's hypothetical organism example in order to make it more concrete, he replaces F by the presence of a lion, R by G(r), s_1 by *roar* and s_2 by *mane*. Loewer's revision does not have any effect on the fundamentals of the example. It just makes the example more

concrete. Using this revised example, Loewer raises two main objections against Dretske's solution for the problem of misrepresentation. The following quote is a good summary of Loewer's first objection.

> We have been assuming that, although various s_i can come to trigger G(r), those which do are all correlated, under normal conditions, with the presence of a lion. Dretske doesn't tell us exactly what constitutes normal conditions. Presumably he would count the presence of holographic images of lions and or tape-recorded roars as not normal. (Emphasis original) [Loewer 1987, p.177]

Instead of pointing to a weakness of Dretske's approach, Loewer's objection shows his misunderstanding of the approach. This approach does not rely on distinguishing normal conditions from abnormal ones; rather, as mentioned above, it relies on the instances of natural meaning and the instances that do not count as natural meaning. At every given instance, G(r) $means_n$ (natural meaning) s_1 or s_2 or $\ldots s_k$ for some specific k. If a lion is present, these s_i also $mean_n$ that a lion is present. This is because of the factivity principle attached to natural meaning. Loewer thinks that these are the normal conditions for Dretske. These have nothing to do with being normal or abnormal. It is the triviality of natural meaning that makes a lion present in these instances. In other instances where a lion is not present, G(r) still $means_n$ s_1 or s_2 or s_k for some specific k, but the natural meaning relationship between these s_i and the presence of lion does not hold anymore. It is worth noting that these are all about natural meaning. Dretske's solution for the problem of misrepresentation is not based on natural meaning. As mentioned in the previous section, Dretske's claim is that natural meaning cannot even be a starting point for the problem of misrepresentation, since natural signs either do their job or do not do it at all. This is why he seeks another type of meaning. This new type of meaning, functional meaning, carries the possibility of not fulfilling its task, in some instances, and thereby the possibility of misrepresentation. Functional meaning is the solution for the problem of misrepresentation, not natural meaning.

Once the function of a mechanism (or a state) of an organism is identified properly, there is no distinction between any conditions. In other words, the distinction between normal versus abnormal conditions does not make any sense after the proper identification of the function of a mechanism or a state of an organism. Using Loewer's version of Dretske's hypothetical organism, if we assume the function of G(r) is identified as indicating the presence of a lion, the following statements will be true at a given time t_1.

A. G(r) means_n s_1 or s_2 or or s_k for a specific k.

B. G(r) means_f that a lion is present.

Now, by a simple application of the law of excluded middle, either a lion is present or it is not. If a lion is present than the following statements in addition to A and B will hold.

C. s_1 or s_2 or or s_k means_n that a lion is present.

D. This is not an instance of misrepresentation rather it is a positive instance of representation.

If *a lion is not present* than the following statements in addition to A and B will hold

C. s_1 or s_2 or ... or s_k does not mean_n that a lion is present.

D. This is an instance of misrepresentation, or at least false content, since B is true and no lion is present. To put it differently, the state of the organism in question is not fulfilling its function properly.

One may claim that identifying the proper function requires a distinction between normal and abnormal conditions; this is different than Loewer's objection although it shares a basic motivation. Since the function of a mechanism or a state of an organism is indeterminate as Dretske himself claims, one may think that the function of a mechanism will be determined on the basis of normal conditions in which the function of the mechanism is fulfilled. In other words, the following simple algorithm could be used for removing the indeterminacy of function: i) find the normal conditions; ii) analyze the behaviors of the mechanism; iii) find the function that is fulfilled across all the instances of normal conditions. If this were to be Dretske's claim, then Loewer's objection would succeed against Dretske's suggestion. As the first step of the simple algorithm clearly suggests, one needs

to be able to have some criteria for defining normal conditions. However, Dretske's suggestion for removing the indeterminacy of function is totally different; it has nothing to with finding normal conditions. His solution for the indeterminacy of function comes from the complexity threshold of an organism. For primitive organisms such as marine bacteria, it is not possible to eliminate the indeterminacy of the functions of their mechanisms. For example, the function of magneto taxis mechanism could both be the proximal stimulus or pointing to comparatively oxygen-free environment. When organisms are more complex, in other words when they pass a complexity level threshold, then it is possible to eliminate indeterminacy because proximal stimuli cannot be the functional meaning of a state (or a mechanism) of an organism. In Dretske's hypothetical organism with an associative learning mechanism, the proximal stimuli that are attached to the functional meaning change continuously since it is possible to condition the organism with new stimuli. This continuous change, as mentioned above, is not consistent with the need of a time-invariant function that must be attached to the state of the organism in question.

If this complexity threshold approach is justified, then Dretske's solution does work. If not, which is the case as we have seen above, then Dretske needs to tell us normal conditions from which one can identify the function of the mechanism. But Loewer's first objection does not start with that point, hence it shows his misunderstanding of Dretske's solution.

Loewer's second objection is more on the right track. Loewer says,

> He [Dretske] has told us that G(r) represents for O that w is F if it is the function of G(r) to carry information that w is F. He does not provide us a positive characterization of 'the function of G(r) is to carry the information that p.' [Loewer 1987, p.177]

The complexity threshold solution for identifying the function of a mechanism is based on an unjustified process of elimination. Dretske claims that the function cannot be to detect proximal stimuli, because that would not lead to a time-invariant function, hence it should be to detect the existence of F (or lion in Loewer's version). So, as Loewer points, it is not a positive characterization. To detect F is the only option for the time-invariant functional

meaning (as the reader might remember this is not a requirement for natural meaning). Dretske states this claim without any justification. I think this is the fundamental flaw in Dretske's reasoning.

Loewer continues his objection by suggesting a positive characterization of representation within Dretske's framework. His suggestion is as follows.

> G(r) represents that w is F iff the most specific information common to every possible token of G(r) which occurs when conditions are normal is that x is F and O needs the information that w is F. [Loewer 1987, p.177]

Although Loewer finds this characterization of Dretske's solution an improvement over the solution found in Dretske's earlier book *Knowledge and Flow Information*, he points out two main problems about it. In order to have a naturalistic account on the basis of the above characterization, he says, one needs to characterize normal conditions and O's informational needs in non-intentional terms. His claim seems a plausible one at first glance. However, once again, the problem for Dretske is to eliminate the indeterminacy of function. Once it is successfully identified, then one does not need to talk about normal conditions nor about the informational needs of the organism.

Although I agree with Loewer that Dretske's solution for the problem of misrepresentation is, at best, problematic, I think he attacks Dretske's solution from the wrong front. Instead of focusing on the criteria for identifying normal conditions, which is related to Dretske's solution indirectly, he should focus on the complexity level criterion for identifying the function of a mechanism (or a state of an organism).

In conclusion, Dretske's solution for the problem of misrepresentation is not satisfactory. That is to say, his theory does not fulfill the requirements of the fifth preliminary discussed in Chapter 1. In this dissertation, I offer a probabilistic theory of mental content that agrees with Dretske's framework with respect to the first four of the preliminaries. Moreover, different than Dretske's theory, it satisfies the fifth preliminary as well. Thus, the theory I offer solves the problem of misrepresentation. Before proceeding with the details of my theory, however, there is another important feature of Dretske's theory that needs to be discussed:

the use of inverse conditional probabilities. In the philosophical literature, Dretske's theory is the first attempt at using inverse conditional probabilities instead of forward conditional probabilities. I believe that such an approach is very useful for understanding mental entities. Therefore, I borrow Dretske's use. However, Dretske ends up reversing the advantages of using inverse conditional probabilities by assigning unity to them in his definition of informational content. I borrow his use of inverse conditional probabilities without making his mistake. In the next chapter, I discuss the use of inverse conditional probabilities within a relatively broad historical and interdisciplinary perspective in order to justify my use of inverse conditional probabilities in Chapters 4 and 6.

CHAPTER 3

Inverse Conditional Probabilities: An Alternative Perspective

There is an apparent difference between the first person perspective and the third person perspective in the study of mental entities. The third person perspective provides a more neutral and scientific framework whereas the first person perspective is more in line with semantic qualities of mental entities. Which one is the right perspective for understanding mental entities is an essential question. Dretske's 1981 framework defines mental content in terms of the informational content of a mental state. Dretske provides a probabilistic definition for the notion of informational content of a signal. The conditional probability that he uses in the definition is an inverse conditional probability, i.e. the probability of the stimulus given the mental representation. This corresponds to the first person perspective. The third person perspective would require a forward conditional probability, i.e. the probability of the mental representation given the external stimulus. Dretske's 1981 framework is the first attempt at using inverse approach in the philosophical literature. I believe that the first person perspective, i.e. inverse conditional probabilities, is the right way of defining mental content. As we will see in the following sections, though, Dretske's theory ends up reversing the benefits of using inverse conditional probabilities because of another characteristic of his notion of informational content. The probabilistic theory that I develop and defend in this dissertation borrows Dretske's motivation of using inverse conditional probabilities without making the mistake that leads to reversing the benefits of using inverse conditional probabilities. The details of my theory will be clarified in Chapter 4 and Chapter 6. In this chapter, I provide arguments in favor of using inverse conditional probabilities in a historical and interdisciplinary context by drawing lessons from Philosophy and Neuroscience. After that I discuss the main objection against using inverse conditional probabilities. The objection

is that none of available interpretations of probability is consistent with inverse conditional probabilities. I discuss this objection and argue against one suggestion made based on this criticism. The suggestion is to use counterfactuals for defining informational content instead of using inverse conditional probabilities. In the end, I show that this suggestion does not provide any advantage and I provide a solution for the probability interpretation. Thus, I defend the use of inverse conditional probabilities. Before diving into these details, though, a terminological preamble is needed, because drawing ideas from two different fields sometimes leads to confusing terminology. First, a terminological preamble ...

1. Terminological Preamble

The main subject of this chapter is the right methodology of studying mental states both philosophically and empirically. The main methodology used for understanding physical masses, planets, molecules, tectonic plates is generally called the third person perspective. This methodology is sometimes called the view from nowhere' [Nagel 1989], that is to say, the scientist aims to have a neutral approach without influencing the entity that s/he studies. In the following pages, this methodology is sometimes called the observer's perspective. In other words, the following two expressions are used interchangeably: the observer's perspective and the third person perspective.[1]

Whether such a methodology is applicable to mental states or not is an essential question that has been discussed more or less since late 18^{th} century. If this question is answered negatively, then one needs to offer an alternative methodology for studying mental states. The alternative methodology rests on the assumption that it is not possible to take a neutral third person perspective for studying mental states. The perspective that is needed is the first person perspective. 18^{th} and 19^{th} century philosophy and psychology proceeded with

[1]In the literature the first person vs. the third person perspective distinction is used in a specific way. This way suggests that the first person perspective has some ontological connotations akin to qualia. I do not want to take any position in regards to existence of qualia. I am using this distinction only because it is familiar to philosophers. The only thing that I need for the purposes of this dissertation is to make a point about being an observer in evaluating mental states or being the evaluator. Prof. Frederick Schmitt pointed out this problem for an earlier draft of this dissertation.

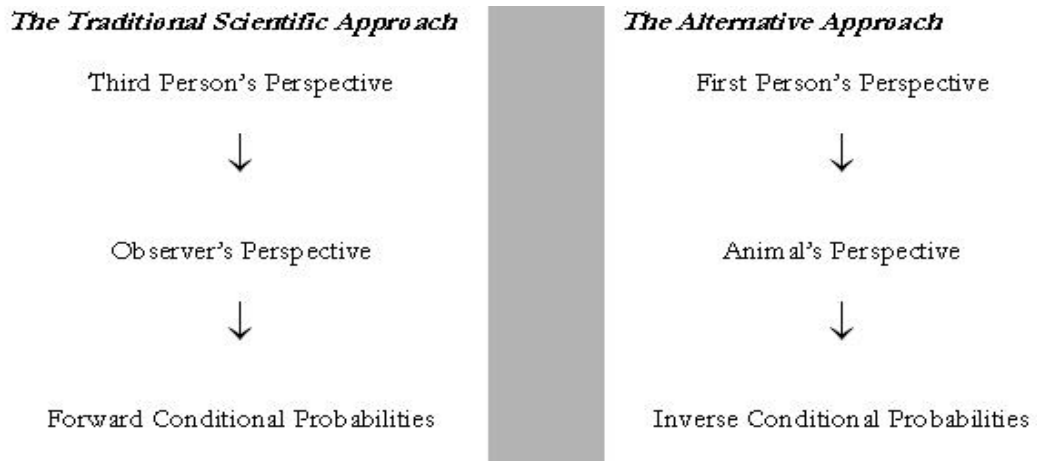

FIGURE 1. Terminology Summary

this methodology and ended up with Wundt's introspectionism. A similar trend is true for history of biology. 19th century biologist, Von Uexkull coined the term 'the animal's perspective' for this purpose [Uexkull 1926]. Following this tradition, 'the animal's perspective' and 'the first person perspective' are used interchangeably in this dissertation.

Lastly, when these two different perspectives is translated into the language of probability calculus, the third person perspective corresponds to the notion of forward conditional probabilities, whereas the first person perspective corresponds to the notion of inverse conditional probabilities. The former means the probability of mental response given an external stimulus - P (r | s) - and the latter means the probability of an external stimulus given a mental response - P (s | r).

Figure 1 summarizes the expressions that are used interchangeably for each methodology.

2. The Traditional Approach: Forward Conditional Probabilities

Whether mental states are proper scientific entities that can be studied empirically has been a philosophically legitimate question since the late 18^{th} century. To answer this question requires comparing mental states with other legitimate scientific entities such as atoms, molecules, cells, quarks, tectonic plates etc. Subject matters of all other empirical

disciplines such as physics and chemistry have one essential feature in common. It is possible to take a neutral third person perspective, that is to say the scientist can take the observer's point of view in order to identify the general features and laws that govern those entities. It is not obvious whether this is true of mental states. This problem is the basis of the debate about whether Psychology is a legitimate empirical science or not. The current debate about the animal's and the observer's perspectives in theories of mental content falls under this broader discussion about the status of mental states and the status of Psychology. Thus, a historical analysis of the broader discussion is valuable for understanding this debate.

2.1. Historical Context. Immanuel Kant's *Metaphysical Foundations of Natural Science* is one of the early works where this question about the status of mental states as scientific entities and the status of Psychology as a legitimate empirical science are discussed in a systematic way. In this treatise, Kant compares Psychology with other empirical sciences, especially with Chemistry, and claims that Psychology is not a legitimate empirical science. He presents three arguments for this claim. His first argument relies on the premise that mental states do not have spatial dimension. Since they don't have spatial dimension, then there is no way of doing real experiments, i.e. mental states cannot be studied as one can study molecules or atoms. His second argument is a corollary of the first one. He says that there is no way of collecting data about mental states prone to mathematical modeling. If it is not possible to do empirical experiments about mental states, then there is no way of collecting data about mental states. The third argument is about the main subject matter of Psychology. The subject matter of psychology, he says, is the self, but it is impossible for the self to study its own workings let alone to do so in a disinterested way. From these three arguments, or rather two arguments and one corollary, he concludes that mental states are not legitimate scientific entities and Psychology is not a proper empirical science [Kant 2004].

Despite the fact that Kant presents two arguments and one corollary, all of these rely on one main idea. The idea is that it is not possible to take 'the observer's perspective' with respect to mental states, and since 'the observer's perspective' is essential for empirical

sciences, mental states are not legitimate scientific entities. This Kantian idea, which could be called the Kantian legacy, dominated the entire 19^{th} century. Many physiologists and philosophers, especially in Germany, tried to defeat the Kantian Legacy and show that it is possible to study mental states empirically. For example, Herbart, a German physiologist, claimed that mental states and ideas exhibit the variables of time, intensity and quality. Hence, it should be possible to measure for acquiring real data. By the end of the 19^{th} century, the legacy was almost defeated. The pioneers of this achievement were Herbart, Helmholtz, Fechner, and Donders. Among these, Helmholtz is the most influential one. In order to show the possibility of empirical methods in studying mental/psychological states, he measured the length of time it takes to transmit impulses along a nerve. He questioned Kant's beliefs about innate ideas of space and used the notion of unconscious inference for explaining visual illusions. Fechner, following Helmholtz's footsteps, proved that not only is it possible to study mental/psychological states from a third person perspective but it is also possible to find laws that govern our mental states. He found the law that governed the relation between the intensity of a sensation and the strength of features of a stimulus. This law, which is still in use, states that the intensity of a sensation given a stimulus is equal to the logarithm of the stimulus's relevant feature.

In short, the efforts of German physiologists and philosophers defeated the Kantian legacy, and showed the possibility of taking the observer's perspective in studying mental, psychological and neural states. As a result, the observer's perspective (the third person perspective) has become the main methodology in Philosophy, Psychology and Neuroscience. Needless to say, this methodology has improved our understanding of mental states by producing useful empirical results. However, because of its very nature, it also limited our understanding of mental states. As mentioned above, this chapter aims to analyze these limits and defend a different methodological perspective which should be used for the endeavor of unveiling the mystery of the human mind. Before doing that, let me show the prevalence of the observer's perspective as the main methodology in Philosophy and Neuroscience as conducted in the 20^{th} century.

2.1.1. *Philosophy.* In philosophical literature, providing a satisfactory theory of mental content is the first step for understanding mental states. As stated in the previous chapters, there are several different approaches for explaining the content of mental states, but this dissertation limits itself to causal/informational theories. So, I shall discuss the dominance of the observer's perspective in this type of theories. The main problem that afflicts causal/informational theories is the problem of misrepresentation and its sister problem, the disjunction problem. The way these two problems are phrased clearly carries the effect of the dominant methodology. Suffice it to show that the dominance of the observer's perspective in the formulations of these two problems for the purposes of this chapter.

Rowlands's (1999) description of the problem of misrepresentation is a good example of the way the problem is defined in the literature.

> Consider a mental representation of a horse The representation HORSE, it seems, means 'horse'. This is what makes it the representation it is, and not the representation of something else. However, it also seems possible, indeed likely, that the representation HORSE can be caused by things that are not horses. Donkeys in the distance and cows on a dark night might, in certain circumstances, be equally efficacious in causing a tokening of the HORSE representation. Now, according to an informational [or causal] account, representation is to be explained in terms of nomic dependence. However, if the representation HORSE can be tokened in the absence of horses, then HORSE does not seem nomically dependent on horses in the relevant sense. Rather, what HORSE does seem nomically dependent upon is not the property of being a horse but the disjunctive property of being a horse or a donkey-in-the-distance or a cow-on-a-dark-night. Thus, if information is a matter of nomic dependence, and if representation is a matter of information, then we seem forced to say that what HORSE represents is not the property of being a horse but the above disjunctive property. [Rowlands 1999]

As should be obvious from the above quote, the dependency relation being favored is the dependence of the representation HORSE on the entities that cause (or trigger) an instantiation of the representation HORSE. This approach, analyzing the content of a mental representation given its cause and/or its informational source, is the common practice. This common practice is an instance of taking the observer's perspective (i.e. the third person perspective), which is then translated into the language of probability theory as forward conditional probabilities.

The disjunction problem and the problem of misrepresentation are two sides of the same coin. So, it should not come as a surprise to find out that the same perspective is used for describing the disjunction problem. Jerry Fodor, who originally formulated the problem and named it in 1984, also uses the same perspective in his original formulation. Not only does he take the observer's perspective in the formulation, but also clearly shows his aversion to the other alternative, i.e. the reverse approach which is the animal's perspective. His aversion becomes crystal clear when he discusses Stampe's epistemic suggestion for determining the content of a mental representation (Stampe's epistemic suggestion is one of the early and naturally immature attempts of using the animal's perspective). Fodor says,

> Now, generally speaking, if representation requires that S cause R, then it will of course be possible to learn about S; inferring from their effects is a standard way of coming to know about causes. So, depending on the details, it's likely that an epistemic account of representation will be satisfied whenever a causal one is. But *there is no reason to suppose that the reverse inference holds.*(emphasis is mine) [Fodor 1992]

Another example of the dominance of the observer's perspective (and forward conditional probabilities) is found in Dretske's 1988 book, *Explaining Behavior*. In this book, Dretske categorizes representational systems into three: Type I, Type II and Type III. Only Type III systems have full representational and intentional capacity (the other details of these systems are not essential for the current discussion). In order to explain the representational capacity of the third type, Dretske appeals to neuroscience. The specific

experiment that he cites is Lettvin's 'bug detector' experiments in a frog's visual field. As will be discussed in detail in the following section, this experiment uses the conditional probability of the neural response given a bug. Hence, it also takes the observer's perspective [Dretske 1988]. Dretske's appeal to this specific experiment shows his inclination toward the commonly accepted observer's approach. As a side remark, it should be noted that this is not the whole story about Dretske's theory. In his earlier book, Dretske adopts the animal's perspective instead of the observer's perspective. This point will be discussed in the following sections.

2.1.2. *Neuroscience.* Rieke et al. in their book, *Spikes: Exploring the Neural Code*, [Rieke *et al.* 1999] give a quick overview of the history of neuroscience with the aim of discussing the benefits of the animal's perspective over the observer's perspective. As they claim, E.D. Adrian's experiments, conducted roughly between 1915 and 1925, established three fundamental facts about the neural code. These facts determine the methodology in neuroscience to this day. Adrian's three facts are as follows.

- All or none law: neural activations are all or none, that is to say, individual sensory neurons produce stereotyped action potentials (spikes)
- Rate Coding: in response to a static stimulus such as a continuous load on a stretch receptor, the rate of spiking increases as the stimulus becomes larger
- Adaptation: if a static stimulus is continued for a very long time, the spike rate begins to decline.

Based on these facts, Adrian measured the response of a neuron by counting the number of spikes in a fixed time after the onset of the stimulus. This clearly falls under the category of the observer's perspective and forward conditional probabilities. Following Adrian's legacy, the typical approach in modern experiments is to repeat the same stimulus many times and average over these repeated presentations in order to find the number of spikes when the neuron in question is presented with a stimulus. The typical approach based on the observer's perspective has been in used in neuroscience since the time of Hubel and

Wiesel's classic study on the cat's visual cortex. In this study, these two researchers implanted a microelectrode into the visual cortex of an anaesthetized cat. Then, they recorded activations in the cells in response to different patterns of light. The experiment showed that some cortical cells are activated more strongly when presented particular patterns of light [Hubel & Wiesel 1962]. The measurement method in the study is to measure the activations of the cortical cells given a specific stimulus. The methodology, hence, is the observer's perspective and forward conditional probabilities. Another example of the same methodology is Lettvin's experiment about the frog's retinal ganglion cells [Lettvin *et al.* 1940]. Lettvin, after presenting several bugs to the frog's visual field, found that retinal ganglion cells are activated whenever presented by a similar visual stimulus. Hence, these ganglion cells are called 'bug detectors' (whether they really detect bugs or not has initiated a lively philosophical discussion).

3. The Alternative Approach: Inverse Conditional Probabilities

Despite the dominance of the observer's perspective in both Philosophy and Neuroscience, there have been notable attempts at using the alternative approach, the animal's perspective, as well. Such attempts in Neuroscience are more deliberate than the ones in Philosophy. Most of the attempts in Philosophy stemmed from concerns other than methodological ones. However, the attempts in Neuroscience have resulted directly from methodological concerns. Let me start summarizing these attempts in these two fields, starting with Neuroscience.

3.1. Neuroscience. Fitzhugh's analysis of optic nerve messages (1958) is the first attempt to use the animal's perspective as opposed to the observer's perspective. In his analysis, he proposes 'a statistical mathematical model of the discharge in a single optic nerve fiber.' For his model, he uses empirical results from the activations in retinal ganglion cells of a cat in response to a flash of a light. His very first observation regarding the mammalian optic nerve is that the activations are noisy. The general way of dealing with noisy signals in communication theory is to divide the incoming signal into two, stimulus and noise, and analyze the magnitude of the incoming signal as the algebraic sum of the

stimulus and the noise. This linear summation assumption, says Fitzhugh, is not physiologically plausible for the mammalian optic nerve. That is to say, there is no way of separating the incoming signal into two separate categories as stimulus and noise. This is his main rationale for proposing a non-linear stochastic model of analyzing the mammalian optic nerve message. In his words,"It is necessary, therefore, to set up a non-linear stochastic model of the nerve discharge and apply to it those more general parts of communication theory which deal with the process of statistical inference, based on the calculation of likelihoods. [Fitzhugh 1958, p.676]" The calculation of likelihoods is the key notion that shows taking the animal's perspective and using inverse conditional probabilities. His statistical analyzer modeled on the use of inverse conditional probabilities successfully accounts for 'the detection of the occurrence of a flash of light of known intensity and time of occurrence, the detection of the time of occurrence of a flash of known intensity, and the estimation of the intensity of a flash occurring at a known time.'[2]

In short, Fitzhugh's work shows the value of using the probability of the stimulus given the response, P (s | r), instead of using the probability of the response given the stimulus, P (r | s). The former being the animal's perspective and the latter the observer's perspective.

[2]The other feature of Fitzhugh's statistical analyzer, which is relevant for the purposes of this dissertation, is the frequency function that he uses in the filter that sums up the optic nerve message that spans over 100 msec into a signal at a single instant of time. This frequency function is determined by the causal interactions but is autonomous than the causal interaction level. This is very reminiscent of the two level analysis that I propose in the fourth and sixth chapters of this dissertation. Here is how Fitzhugh explains his frequency function. 'The use of a frequency function to represent the distortion of the time scale of the maintained discharge was found to be the simplest way to introduce the effect of a stimulus, but this 'time-distortion process' contains some inherent assumptions which may not correspond to physiological reality. Since the frequency function f(t) is determined by the stimulus and in turn determines the statistical properties of the discharge, it forms an intermediate step in the model, separating a causally determined process from a statistical one. In the first process the message is coded without loss of information into a form suitable for neural transmission, namely f(t), the running average frequency of impulses, while the second introduces, in the impulse intervals, the statistical variations which destroy information, However, it is not claimed that these two stages correspond to separate physical processes in the retina; they have been separated only to simplify the theoretical analysis'.[p.687]

Since Fitzhugh's proposed model is a mathematical model, it is only normal to examine the mathematical relation between these two types of probabilities. In fact, mathematically each of these can be converted to the other by a simple algebraic manipulation. According to the Bayes' rule,

Rule 1. $P(A|B) = \frac{P(A\&B)}{P(B)}$

One gets the following equations when Bayes' rule is applied to both inverse and forward conditional probabilities.

Eq.1. $P(r|s) = \frac{P(r\&s)}{P(s)}, thus P(r|s) \times P(s) = P(r\&s)$

Eq.2. $P(s|r) = \frac{P(r\&B)}{P(r)}, thus P(s|r) \times P(r) = P(r\&s)$

These two lead to the following equation.

Eq.3. $P(r|s) \times P(s) = P(s|r) \times P(r)$

In other words, these two types of probabilities can be converted to each other if P(s) and P(r) are both known. One useful metaphor for understanding the relation is the metaphor of dictionaries. Responses and stimuli can be considered as analogues of languages - say for example, Turkish and English. Then, the inverse conditional probabilities would correspond to a dictionary from Turkish to English and the forward conditional probabilities would correspond to a dictionary from English to Turkish. Ideally, if a complete dictionary of Turkish words in English existed, then it would be possible to construct a complete dictionary of English words in Turkish[3]. This, in principle, is possible. The equation that shows conversion between P(r | s) and P(s | r), i.e. Eq3, is similar to this case. And a complete dictionary corresponds to knowing both P(r) and P(s) in order to calculate the other type of conditional probability. P(r) is the probability of the specific response in the organism's mental life and P(s) corresponds to the probability of the given stimulus in the natural order of the world. Although it is very difficult to know P(r), it is not impossible. On the other hand, to know the latter, i.e. P(s), is impossible since it requires a complete

[3]This claim assumes that both Turkish and English have the same expressice power. This may not be true. If it turns out that these two natural languages have different expressive powers, I could run the same argument with some other two natural languages which have similar expressive power. I thank Prof. Frederick Schmitt for raising this point.

past, present and future picture of the natural order in the world. The dictionary metaphor explains this successfully. Despite the fact that it is theoretically possible to construct a complete English to Turkish dictionary from a full fledged complete Turkish to English dictionary, such a goal is unattainable in practice. In short, the mathematical equality stated in Eq.3 does not import any substantial idea regarding the equal status of forward and inverse conditional probabilities (the observer's perspective and the animal's perspective). Quite the opposite, the analysis of the Eq.3 is evidence for why the animal's perspective (inverse conditional probabilities) should be preferred over the observer's perspective (forward conditional probabilities). Inverse conditional probabilities require knowing the joint probability, Pr (r and s), and the probability of the response, Pr (r). Both of these are available within the organism[4] Although to know Pr (r), as mentioned above is practically very difficult, it is not impossible. However, forward conditional probabilities require knowing the joint probability, Pr (r and s), and the probability of the stimulus, Pr (s). It is easy to know the probability of a stimulus in a laboratory environment since it is artificially determined by the scientist, but to know the probability of a specific stimulus occurring in the world is impossible (excluding the possibility of the scientist's being as omniscient as God). Hence, inverse conditional probabilities are preferable over forward conditional probabilities despite the equality that follows from the Bayesian rule.

Eliasmith, probably the first philosopher who used the term 'the animal's perspective' in the mental content literature, discusses the value of the animal's perspective quite extensively in his unpublished dissertation [Eliasmith 2000]. He claims that the equality stated in Eq.3 shows that both inverse and forward conditional probabilities are equally valuable. In a nutshell, his argument goes as follows. The most important thing there is to know for a philosopher of mind or for a cognitive neuroscientist is the joint probabilities for the entire set, i.e. P(r and s). Once this is known, he continues, then one can get either of the

[4]As pointed out by Prof. Frederick Schmitt the term 'available' is a little bit vague in this context.What I mean by it is the following: the organism should be able to form the required probabilities either via its instantaneous interaction with the external world or via its past history. If the probabilities require anything other than either of these two alternatives, then those probabilities are not available to the organism.

conditional probabilities easily. This argument is flawed, because of the reasoning that I presented in the previous paragraph. His very first premise is up to debate. It is not at all clear that the most important value is the values of the joint probability, but even granting this assumption to him does not make his argument work. As mentioned above, forward conditional probabilities require the probability of a given stimulus in the world together with the joint probability function, whereas inverse conditional probabilities require the probability of a given response within the organism's system together with the joint probability function. The organism has access to the probability of his system's response just by simply looking at his past, but it does not have access to the probability of a specific stimulus in the natural order of the world. Hence, Eliasmith's argument from the joint probability function does not have merit.

The other advantage of using inverse conditional probabilities (i.e. taking the animal's perspective) is that it reduces the complexity of the task of constructing the original signal sent from the received message. As stated above, Fitzhugh mentions the non-linear relation between the input and the noise in the case of messages received by the mammalian optic nerve. It is natural to assume that such a non-linear relationship holds for every message that sensory mechanisms receive from the world. Due to the non-linearity, says Fitzhugh, classical solutions of the communication theory for separating the noise from the real message do not work for sensory mechanisms. Rieke et al analyze the non-linearity characteristic of neural activations under two possible perspectives: the observer's and the animal's perspectives. They conclude that adopting the animal's perspective, i.e. using inverse conditional probabilities, produces a more or less linear function by which the external stimuli could be predicted. Contrarily, using forward conditional probabilities does not change anything about the non-linear characteristic of neural activations. In other words, using inverse conditional probabilities reduces the non-linearity of the problem to a linear one. They illustrate their conclusion with an example from an experiment on a motion sensitive cell (H_1) in the blowfly.

In the experiment 'the fly viewed a spatial pattern displayed on an oscilloscope screen, and this pattern moved randomly, diffusing across the screen. At the same time, spikes

[responses of the fly] were recorded.' [Rieke *et al.* 1999, p.26] Figure 2 shows their analysis of the data from two different perspectives. On the left hand column is the inverse description; on the right hand column is the forward description. P (n) is the probability of finding n responses in a fixed time window (Part (b) of the figure). V is the value of the stimulus velocity averaged over the same time-window and P (v) (Part (a) of the figure) is the probability density for all time windows in the experiment. P (n,v) is the joint probability density, i.e. the probability of both the presentation of the stimuli and the response occurring together. And it is depicted in the part (c) of the figure.

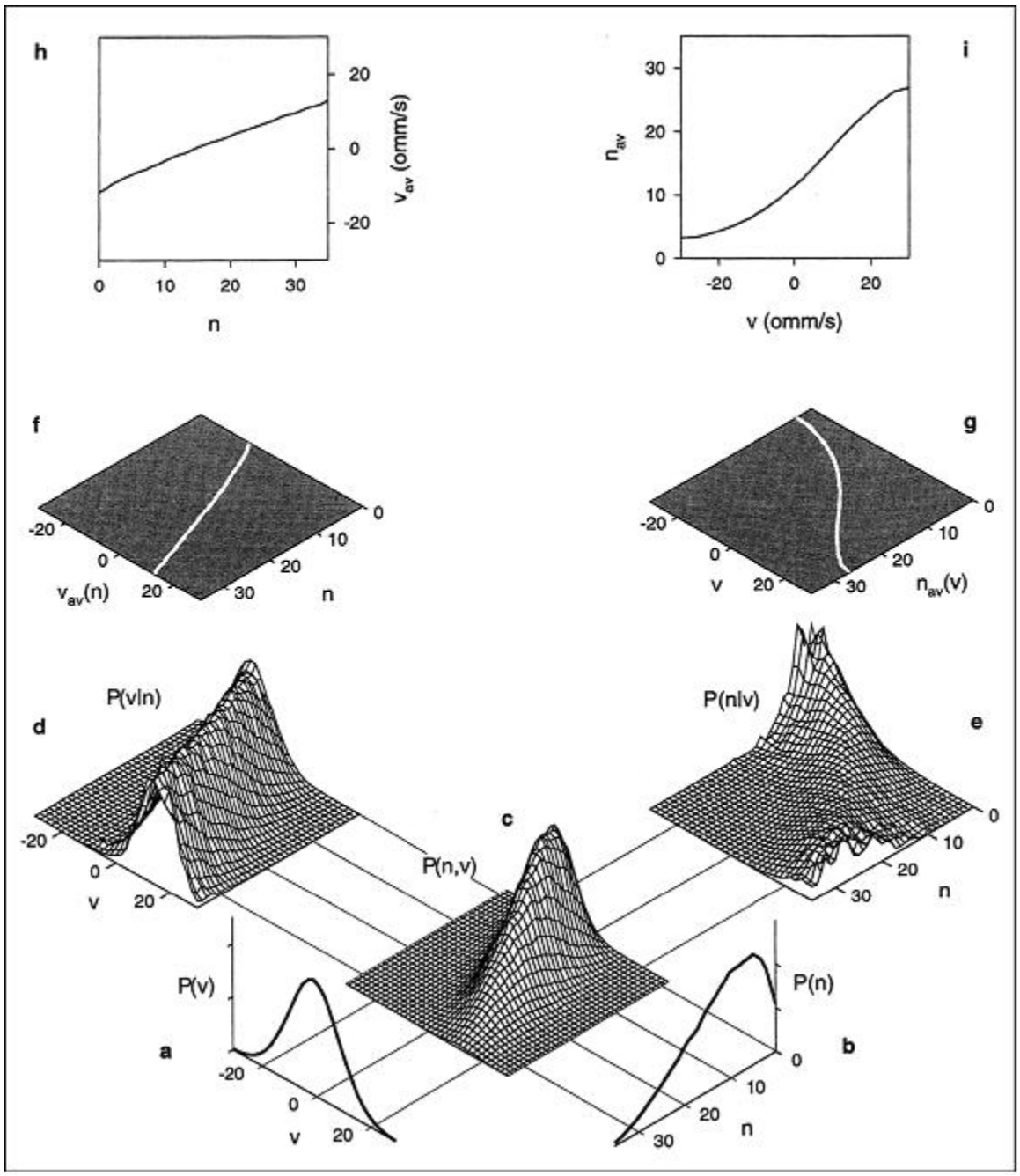

FIGURE 2. Two Different Analyses

The inverse description amounts to asking the likelihood of the stimulus being present for each particular n and it is summarized in P (v | n) shown in part (d) of the figure.

The task in the forward description is to find out what values of n could be induced by a given value of v. This is the conditional probability of P (n | v) as shown in (e) and 'the white lines in panels (f) and (g) show the average values of v given n, and of n given v respectively. These data are replotted in a standard orientation in (h) and (i).' So, the panel (h) is the result of inverse analysis. It is very close to being a line, therefore shows a linear relation between n and v. It is not identical to a line, but it is as linear as one can get in any empirical field. On the contrary, the result from the forward analysis presents a sigmoidal function which shows a non-linear relation between analyzed v and n. Rieke et al. describes the differences between these two results in the following way.

> For the distribution P (n | v), this mean is the average number of spikes produced as a function of the stimulus amplitude, and this is the traditional measure of neural response. We see the familiar non-linear, sigmoidal relationship observed by Adrian that has been reproduced in many systems. But when we ask for the mean of the distribution P (v | n), which is the average stimulus velocity given that we observe a particular spike count, we see a very different, almost perfectly linear relation. The nonlinearity of the sigmoidal input/output relation, so ubiquitous in neurobiology, seems to have vanished. [Rieke *et al.* 1999, p.28]

In short, in neuroscientific literature, the dominant and traditional approach has been using the observer's perspective. In more technical terms, as put by Rieke et al, 'the traditional approach to studying the neural code has been to catalogue the average behavior of neurons in response to changes in stimulus parameters' [Rieke *et al.* 1999, p.60] However, the organism is not interested in predicting responses from known stimuli, because the organism already knows these responses; its task is to predict stimuli from known responses. Moreover, as discussed above, this inverse approach has significant advantages over the traditional forward approach.

3.2. Philosophy. The situation in philosophical literature in terms of adopting the animal's perspective is not even as promising as it is in neuroscientific literature. The

first signs of adopting such an approach could be found in Dennett's book *Consciousness Explained*. In the early parts of his book, Dennett compares the status of observers and the brain in terms of evaluating neuronal activations and realizes that the brain is blind to external conditions that produce those neuronal activations.

> Whereas we, as whole human observers, can sometimes see what stimulus conditions cause a particular input or afferent neuron to fire, and hence can determine, if we are clever, its 'significance' to the brain, the brain is 'blind' to the external conditions producing its input and must have some other way of discriminating by significance. [Dennett 1969, p.48]

Another passage that favors the animal's perspective in Dennett's book is the following.

> [T]he investigators working with fibers in the optic nerves of frogs and cats are able to report that particular neurons serve to report convexity, moving edges, or small, dark, moving objects because these neurons fire normally only if there is such a pattern on the retina [Dennett 1969, p.76]

However, Dennett's early realization of the advantage and necessity of taking the animal's perspective did not lead him to develop a theory of content on the basis of the animal's perspective. Despite this, as being the earliest contemporary philosophical work that mentions the animal's perspective, Dennett's work is a valuable first step.

A deliberate mention of inverse conditional probabilities (i.e. adopting the animal's perspective) in philosophical literature is Dretske's 1981 work. In this work, as mentioned in the previous chapters, he tries to explain sensation, belief and knowledge by using the notion of informational content. In his definition of this notion, he clearly uses inverse conditional probabilities. He says that a signal (in our case a mental representation and/or a vector of neuronal activations) carries the information that the source (in our case the world) is in a specific configuration if the conditional probability of that specific configuration in the source given the signal is one. Dretske uses inverse conditional probabilities, and hence adopts the animal's perspective. This is a significant improvement over Dennett's ambiguous comment about the animal's perspective.

Not only does Dretske deliberately adopt the animal's perspective, but also he makes that perspective the basis of his theory of mental content and epistemology. Given Dretske's attempt, one is inclined to think that the situation in philosophical literature in terms of adopting the animal's perspective is pretty good. This, however, is not the case. The reason for this not being the case is no one other than Dretske himself. The strict constraint that he puts on the conditional probability in his definition of informational content, i.e. assigning unity to the conditional probability, reverses all of his effort for adopting the animal's perspective. Such a strict constraint takes us back to the observer's perspective, even worse it takes us to an ideal version of the observer's perspective. In the case of the traditional observer's perspective, the observer has access to some information that is not available to the animal itself such as stimulus conditions that cause a particular input or afferent neuron to fire. Dretske in his definition goes farther and makes the observer infallible. The observer has perfect knowledge of stimulus conditions, because the conditional probability is one.

In short, although Dretske's use of inverse conditional probabilities is a significant improvement over Dennett's ambiguous comments, Dretske ends up sawing the branch that he is sitting on by assigning unity to the conditional probability in his definition of informational content.

After Dretske's 1980 effort, using inverse conditional probabilities (i.e. adopting the animal's perspective) for analyzing philosophical problems about mental content and the mind fell out of favor until recently. One would expect that some would have tried to adopt the inverse approach by revising Dretske's strict constraint about conditional probabilities, especially given the advantage of the inverse approach in neuroscientific research. On the contrary, there was no such effort until 2000[5]. The main reason for this lack of interest is the very notion of inverse conditional probabilities. Loewer (1982) in his BBS commentary on Dretske's book questions the legitimacy of inverse conditional probabilities with respect

[5]Eliasmith finished his dissertation in 2000. His dissertation mainly tries to construct a theory of mental content by adopting the animal's perspective with some limitations. Please see Section 3.1. for these limitations

to plausible interpretations of probability. He rightly claims that if there is no plausible interpretation of probability that explains the notion of inverse conditional probabilities, then the notion should not be used in the definition of informational content. He briefly surveys the available theories of probability, and then concludes that it is not possible to make sense of inverse conditional probabilities in any available interpretations of probability. Moreover, he suggests a new definition for informational content by using Lewis's backtracking conditionals. Loewer's objection turns out to be a strong one. At least, no one has been able to refute his objection so far. Moreover, Dretske's contradictory remarks regarding the theory of probability that he uses for grounding inverse conditional probabilities contributed to the difficulty of overcoming Loewer's objection. However, I do believe that Loewer's objection can be refuted and a plausible interpretation of probability can be given for justifying inverse conditional probabilities and thereby adopting the animal's perspective. In the following section, I discuss Loewer's objection, Dretske's contradictory remarks and the possible ways of overcoming Loewer's objection in detail.

4. The Main Problem of the Alternative Approach

One of the main aims of introducing the concept of probability has been solving Hume's problem of induction. Whether the concept has accomplished this goal or not is not very crucial for the current discussion. What is important here is that the concept of probability led to the probability calculus as a formal system. This is an important achievement. However, the probability calculus is nothing more than a formal device unless an interpretation of the primitive terms of the probability calculus is provided. Several different interpretations of the probability calculus have been offered. The oldest and best known interpretation defines the probability 'as the ratio of favorable to equally possible cases.' By now, it is commonly accepted that this interpretation is either circular or leads to contradictory results. The failure of the classical approach led to several other interpretations. Currently, there are three interpretations (theories) of probability available to use: probability as a degree of belief, probability as a propensity measure and probability as the relative frequency of a series. Loewer in his objection to using inverse conditional probabilities briefly surveys

these three interpretations and claims that the notion of inverse conditional probabilities cannot be accommodated in any of these without unacceptable consequences.

4.1. Theory 1 - Degree of Belief. In this interpretation, probability is simply a measure of degree of belief. If I believe with complete certitude that the sun will rise tomorrow, then obviously the degree of my belief is one and so is the probability of the sun rising tomorrow for me. If I am convinced that flipping of a coin will result with heads as often as it will result in not heads, then the probability of heads is one and a half for me. This definition allows for different probability assignments to the same event. For example, if you do not believe with complete certitude that the sun will rise tomorrow (because of some scientific or superstitious reason), then the probability of that event for you will be different from the probability of the event for me. Hence, this interpretation is completely subjective. There are several problems with this approach. First, degrees of belief do not always coincide with the calculated values of probability. Salmon [1966] gives a very good example of such a discrepancy. Some seventeenth century gamblers believed that the probability of getting a 6 out of four throws of a die was above 0.5, and their belief coincided with the calculated value of that probability. However, they also believed that (on the basis of the previous belief) getting a double six out of 24 four throws of two dice was above 0.5. Those who placed their bets accordingly ended up losing money. One famous gambler, Chevalier de Mere approached Pascal with this question and Pascal calculated that in order to get a probability above 0.5 the number of throws must be 25 not 24. In other words, the calculated value did not coincide with the degree of belief some gamblers of the seventeenth century had. This problem is a direct result of the subjective character of this interpretation.

The second problem with this interpretation is that it does not give a naturalistic basis for the notion of probability. Using this interpretation for explaining the notion of inverse conditional probabilities leads to a non-naturalistic theory of mental content. This result is not acceptable since naturalism is one of the main requirements of a successful theory of mental content - at least in current philosophy of mind. This problem is the reason that

Loewer presents for rejecting the degree of belief interpretation as a possible interpretation of probability for making sense of inverse conditional probabilities. There is not much to do but agree with Loewer. Dretske also accepts that the subjectivist interpretation is not the right one for his interpreting of probabilities that he uses in his definition of informational content.

4.2. Theory 2- Propensity Account. Probability as the propensity of one event causing another event provides an objective interpretation, hence avoids implausible consequences of a subjectivist interpretation. In this interpretation, also known as the dispositional approach, to assign a probability value to an event requires knowing the propensities of the cause of that event. This approach has been favored mainly because of its solution to a problem that afflicts other objectivist interpretations. The problem is the problem of assigning a probability value to single events. As we will see in the following section, the other type of objectivist interpretations (the frequency approach) does not seem to have a satisfactory account of single events. Single event probabilities are accounted for in the propensity interpretation as a dispositional tendency to produce some outcome in a single trial. Besides being able to account for single event probabilities, the propensity interpretation provides an objective basis for probabilities. Hence, the fallacy of some seventeenth century gamblers is also avoided.

Despite its powerful explanations, as Loewer claims, the propensity interpretation is not helpful for inverse conditional probabilities. The propensity account requires a forward definition. An external state of affairs has a propensity of triggering a mental representation. Hence, the probability of a mental representation (response) given an external state of affairs (stimulus), P(r | s), makes perfect sense. However, speaking of the propensity of a mental representation producing an external state of affairs related to that mental representation is not meaningful in the propensity interpretation. Thus, inverse conditional probabilities, the probability of a stimulus given a response, are meaningless in the propensity account. The propensity approach cannot help us either in making sense of adopting the animal's

perspective in terms of inverse conditional probabilities. Let me restate this point with a quote from Loewer himself.

> The propensity of a chance setup C producing outcome e is usually explained . . . as a measure of the causal tendency of C producing [e]. But Dretske is after the converse probability, the probability that r was produced by a chance setup C. This probability is usually not meaningful on a propensity (or for that matter a frequency) interpretation. The point is that P(r | C) may be meaningful but not P(C | r), since there may be no propensity P(C). [Loewer 1983, p.76]

Once again, I think, Loewer is right.

4.3. Theory 3 - Relative Frequency Approach. The other remaining option for interpreting the notion of probability for grounding inverse conditional probabilities is the relative frequency approach. This approach identifies probabilities with relative frequencies in long series of independent repetitions of the same event. To give a simple example, the probability of getting a five in a die throw is identified in the following way. Let's assume that a die is thrown infinitely many times. This will give us an infinite sequence of numbers between 1 and 6.

1, 1, 3, 4, 6, 4, 5, 6, 6, 5,. . .

The ratio of the times that 6 occurs in the sequence to the entire sequence approaches to 1/6. This limit is the probability of getting a five in a die throw. There are two important assumptions underlying this approach. First, each throw is independent from each other and identically distributed. The second assumption is that the sequence in principle is infinitely long. Because of this last assumption, probabilities in this approach are construed as properties of such repeated experiments [or trials], or generally, of mass phenomena, and not of single events as in the propensity interpretation. This presents a big problem for the frequency approach since one wants to be able to account for probabilities of single events as well. For example, we want to be able to talk about the probability of a woman being elected as the president of the United States in 2008. Since such an event is a single event

(i.e. has never happened before), assigning a probability to it is not allowed in the relative frequency approach.

The other problem with the relative frequency approach is that it relies on an idealization. It requires an infinitely long sequence. As Hajek [Hajek 1997] proved, without such an assumption it would be impossible to have irrational probability numbers. From a mathematical point of view, it is desirable to have irrational probability numbers.

These two problems are the main criticisms of the relative frequency approach. Loewer, especially because of the single event problem, claims that the relative frequency approach is not an option for grounding the use of inverse conditional probabilities. In fact, the relative frequency interpretation would not be satisfactory even for forward conditional probabilities given the problem of single event probabilities.

Besides Loewer's criticism, there is another reason for not considering the relative frequency approach as a viable interpretation of inverse conditional probabilities. Dretske, as the first philosopher who used inverse conditional probabilities, seems to be refuting the relative frequency approach. As proved by Carnap in his [Carnap & Jeffrey 1971], the relative frequency approach implies that the probability of one does not guarantee that the event must occur. The main reason for that is the notion of limit that is used in the definition of probability. However, Dretske clearly wants that the probability one must guarantee the occurrence of an event. This is how his theory satisfies the intimate connection between the notion of informational content and truth. Such an intimate tie is essential for him since he wants to define knowledge on the basis of informational content and knowledge implies truth. This is the basis of his strict constraint of assigning unity conditional probabilities. Dretske clearly expresses these considerations in a footnote in his *Knowledge and Flow of Information*.

> There are interpretations of probability (the frequency interpretation) in which an event can fail to occur when it has a probability of 1 (or occur when it has a probability of 0), but this is not the way I mean to be using probability in this definition. A conditional probability of 1 between r and

> s is a way of describing a lawful (exceptionless) dependence between events of this sort, and it is for this reason that I say (in the text) that if the conditional probability of s's being F (given r) is 1, then s is F. [Dretske 1981, p.245]

These remarks are understandable given Dretske's overall project. Dretske does not want to use the relative frequency approach, but he does not offer any interpretation of probability in his book either. Unfortunately, for Dretske, the story of a viable interpretation of probability does not end with the quote above. When criticized by Loewer (and by many others) in 1983, he responds to these criticisms in a way that suggests favoring the relative frequency approach. Here is Dretske's response in his own words.

> I see no reason not to think of these probabilities as relative frequencies (among condition types). But two points must be remembered. First, no finite sample is needed to reflect the actual probability. This, of course, is why frequentists need the notion of a limiting relative frequency - to maintain the distinction between accidental correlations and genuine probabilities. Second, the relationship on which the communication of content depends is the lawful dependence that precludes the occurrence of one without the other. In other words, I do not want to identify, as Lehrer & Cohen suggest I must, a probability of 1 with a limit of 1 (see Chapter 3 n.1 [the footnote quoted above]) [Dretske 1983]

There is an apparent tension between the footnote quoted above and this last quote, to say the least. Dretske does not have any problem with the relative frequency approach as long as a probability of one is not equated with a limit of 1. In other words, Dretske favors a 'revised version' of the relative frequency approach without providing any substantial idea about how such a revision should be formulated.

In short, it seems that none of the available interpretations of probability supports using inverse conditional probabilities for defining informational content and thereby mental content. Given this fact, there are two options for saving a causal/informational theoretic

approach that uses inverse conditional probabilities. The first is to give up inverse conditional probabilities and find a new way of defining informational content. The second option is to be more conservative in terms of using inverse conditional probabilities. That is to say, to keep inverse conditional probabilities by defending the relative frequency interpretation of probability since it is the most promising interpretation among the available interpretations. The former option is suggested by Loewer (1982) and Cohen & Meskin (2006). They offer a new definition for informational content in terms of backtracking conditionals and counterfactuals respectively. Their suggestions, which are very close cousins of each other, are compelling as will be discussed in the following section. However, they are not strong enough to give up the advantages of using inverse conditional probabilities. Hence, I defend the second option. I claim that it is possible to account for single event probabilities within the relative frequency approach by following Reichenbach and Salmon's notion of weight. If I am right, then the best course of action is to keep inverse conditional probabilities in the definition of informational content because of the several advantages that are examined in the previous sections. I discuss the details of my suggestion in the following section after arguing against the Loewer and Cohen & Meskin line of suggestion.

5. The Main Problem: One Expensive Suggestion and The Solution

In their 2006 paper, *An Objective Counterfactual Theory of Information*, Cohen & Meskin argue against the use of inverse conditional probabilities in defining informational content and they suggest a new definition based on counterfactuals. They claim that the definition that they suggest fares better than the general tendency of using inverse conditional probabilities for the definition. Their plan of attack, which is originally provided by Loewer in 1983, has two main steps.

Step 1: Show that the standard accounts of information in circulation use inverse conditional probabilities.

Step 2: show that it is difficult to make sense of inverse conditional probabilities on any of the usual interpretations of probability.

They pick Dretske's theory as the paradigmatic example of standard accounts of information for their discussion. This is a very reasonable choice since Dretske's theory is the most influential and typical one. Cohen & Meskin cite Loewer's criticisms that seem to show none of the available interpretations of probability grounds the inverse conditional probabilities that Dretske's theory needs. I discussed this problem in the previous section in detail. As I claimed without presenting any argument, probability interpretation presents a big problem for Dretske's theory. However, as I will show in the section after this one, this problem is not insurmountable.

As mentioned, Cohen & Meskin choose Dretske's theory as the paradigmatic example with the intention of generalizing their results to other standard accounts of information especially to Shannon & Weaver's mathematical theory of communication. They claim that the probability interpretation problem is applicable also to Shannon's mathematical theory of communication since Shannon's fundamental notion of mutual information makes 'ineliminable' (their word choice) reference to inverse conditional probabilities. Well, they are wrong! First, here is what they say.

> The remarks that follow are applicable to other accounts of information (both semantic and quantitative) that are grounded in conditional probabilities. Most saliently, consider the setup of [Shannon, 1948]: let $\{s_1, \ldots, s_n\}$ be discrete alternative states of a source s with probabilities $\{\mathrm{P}(s_1), \ldots, \mathrm{P}(s_n)\}$ respectively, and let $\{r_1, \ldots, r_k\}$ be discrete alternative states of a receiver r with probabilities $\{\mathrm{P}(r_1), \ldots, \mathrm{P}(r_k)\}$ respectively; assume that $P(s_i) > 0$ for all i, that $P(r_j) > 0$ for all j, and that $\sum_{i=1}^{n} P(s_i) = \sum_{j=1}^{k} P(r_k) = 1$. Shannon defines *the mutual information between s and r* as follows: $I(s,r) = -\sum_{i=1}^{n} P(s_i) \log_2 P(s_i) + \sum_{j=1}^{n} P(s_i|r_j) \log_2 P(s_i|r_j)$. So defined, mutual

> information makes ineliminable reference to the same sorts of inverse conditional probabilities as Dretske's theory, and so is vulnerable to the concerns we raise about the interpretation of those probabilities. [Cohen & Meskin 2006, footnote 1]

P (s | r) in the quote refers to the probability of a stimulus given a response, i.e. the probability of an external state of affairs given a mental representation. Since such a probability proceeds from 'inside' to 'outside', it is an 'inverse conditional probability'. Therefore, Cohen & Meskin reason, the probability interpretation problem applies to Shannon's mathematical theory of communication as well. What they miss, however, is that mutual information is commutative. That is to say, I (s,r) is equal to I(r,s). Hence, the formula for I(s,r), the mutual information between s an r, can be rewritten as the following.

$I(s,r) = I(r,s) = -\sum_{j=1}^{n} P(r_j) \log_2 P(r_j) + \sum_{i=1}^{n} P(r_j|s_i) \log_2 P(r_j|s_i)$.

In this formula, there is no reference to inverse conditional probabilities. It is completely written in forward conditional probabilities and once it is written in that way, as Cohen & Meskin would agree, one can adopt the propensity interpretation for forward conditional probabilities with no problem. Although it is true that Shannon himself favored the relative frequency approach as the probability interpretation, his theory presents no reason for not using the propensity interpretation. The point here is not which interpretation of probability is correct; rather what Cohen & Meskin claim about Shannon's theory is not accurate. The probability interpretation problem does apply to Dretske's theory, but not to Shannon's mathematical theory of communication. Such a mistake, however, does not affect the core of Cohen & Meskin's paper. It just limits the applicability of their criticisms. They could have just focused on Dretske's theory without generalizing their claims to other accounts of information. The counterfactual definition account that they offer is a very valuable contribution to the literature and it needs to be discussed.

For the counterfactual account that they propose, Cohen & Meskin start with a crude counterfactual account. Then, they revise this crude account by adding a non-vacuousness clause because they claim that the crude account leads to some unacceptable consequences

about necessary truths. For each of these, the crude and the revised accounts, they present one weak and one strong version of their account. The main difference between the weak and the strong versions is that the counterfactual criterion is only a sufficient condition for the former whereas it is both a necessary and a sufficient condition for the latter. This difference between their strong and weak versions does not have any substantial effect on their claims about why the counterfactual account should be preferred over the standard Dretskean account. Hence, I use only the weak version as the representative of their position for the sake of simplicity. Here is their weak claim for defining informational content.

> **W:** x's being F carries information about y's being G if the counterfactual conditional 'if y were not G, then x would not have been F' is non-vacuously true. [Cohen & Meskin 2006]

The non-vacuousness clause excludes assigning information carrying relation to cases where y's being G is necessarily true. If y's being G is necessarily true, then the counterfactual will come out true no matter what, hence the counterfactual will be vacuously true. The general intuition about necessary truths is that they don't carry information at all. Following this generally accepted intuition, Cohen & Meskin aim to exclude necessary truths from the set of information carrying signals by adding the non-vacuousness clause.

They claim that the counterfactual based definition for information carrying relation should be preferred over the Dretskean definition based on inverse conditional probabilities. They provide three reasons for this claim. First, the transitivity of information flow, i.e. if A has the information B and B has the information C then A has to have the information C, is a fundamental requirement that needs to be satisfied according to Dretske. This 'intuitive' requirement leads to some unacceptable consequences (these will be discussed in the following section). However, the counterfactual definition does not force the transitivity on information carrying relations, and thus, is immune to the consequences that afflict Dretske's account. Second, they claim that Dretske's account makes essential reference to nomic regularities whereas the counterfactual account is agnostic about whether an appeal to nomic regularities is required for information carrying relations. Thus, they conclude,

the counterfactual definition requires a more economical ontology. The third reason that they provide for why the counterfactual definition should be preferred over the Dretskean definition is about the doxastic states. Dretske's account makes essential reference to background knowledge for information carrying relations. Cohen & Meskin claim that such a reference fails to provide objective, reductive explanations of notions of epistemology and philosophy of mind. On the other hand, their account makes no such reference to background knowledge and has the potential of providing objective and reductive explanations for philosophically interesting notions.

I claim that none of these reasons provides a substantial advantage for the counterfactual account over the Dretskean account. I will present my arguments by analyzing their three reasons one by one.

5.1. Information Flow Properties. Dretske's account defines information carrying relation in the following way.

> **Informational Content:** A signal r carries the information that s is F if and only if the conditional probability of s's being F, given r (and k), is 1 (but given k alone, less than 1) [where k refers to background knowledge]. [Dretske 1981]

The most controversial feature of this definition is assigning unity to the conditional probability. This feature leads to several unacceptable consequences such as denying the possibility of partial information and misinformation. Despite these consequences, Dretske claims that he is obliged to assign unity, because it is the only way to match our common sense intuitions about information flow. Two of these intuitions are what he calls the conjunction principle and the Xerox principle. The former simply claims that If a signal r carries the information that A and if it carries the information that B, then it has to carry the information that A and B. The latter is the transitivity property that is mentioned above: If A has the information that B and B has the information that C, then A has to have the information that C. These are intuitively true claims, and any technical definition of information carrying relation needs to match these two principles. On the other hand,

we know from the basics of the probability calculus, conditional probabilities satisfy these two principles only if they are one (for a detailed explanation of this, please see Chapter 4).

Cohen & Meskin do not discuss the conjunction principle, but they criticize the Xerox principle. They claim that the Xerox principle is true for most cases, but not all cases. That is to say, they claim, the information carrying relation is neither transitive nor intransitive. It is non-transitive. What they claim so far is correct, in fact in the following chapter, I analyze why the Xerox principle does not hold for all cases in detail. I provide both mathematical and philosophical reasons for why the information carrying relation is non-transitive. Cohen & Meskin claim that their counterfactual account matches the limited application of the Xerox principle better than the Dretskean account. In Dretske's framework the following argument is valid.

A has the information that B.

B has the information that C.

Therefore, A has the information that C.

However, as we know from Lewis's possible worlds semantics of counterfactuals, the following inference schema is not valid.

$A\Box \rightarrow B$

$B\Box \rightarrow C$

Therefore, $A\Box \rightarrow C$

This inference is not valid, because the closest possible A world may not be a C world given that the closest possible A world is a B world and the closest possible B-world is a C-world. However, this does not mean that the inference never holds, that is to say, there are cases where the conclusion follows from the premises but not always. It may hold in many cases. The only claim is that there could be cases in which it does not hold. It is true that, as Cohen & Meskin claim, this fits better to the limited application of the Xerox principle. If that were the only feature of the counterfactual account for the intuitive properties of information carrying relation, then it would have constituted as a reason for preferring the counterfactual account over the Dretskean account. However, that is not the

case. The conjunction principle mentioned above provides insurmountable difficulties for the counterfactual account.

There are good reasons for questioning the application domain of the Xerox principle, but it is very difficult to come up with a reason for rejecting the Conjunction Principle. The Conjunction Principle implies that if a signal r carries the information B and if it also carries the information that C, then it has to carry the information B and C. This principle is Dretske's other important reason for assigning unity to conditional probabilities. Cohen & Meskin do not analyze their counterfactual account with respect to the Conjunction Principle. In fact, their counterfactual definition does not satisfy this principle, either. Let's assume that A carries the information that B and A also carries the information that C. According to their definition, A carries the information that B if the counterfactual 'if B were not the case, then A would not have been the case' is non-vacuously true. When this definition is applied to two assumptions, one gets the following counterfactual claims:

(1) 'If B were not the case, then A would not have been the case' is true.

(2) 'If C were not the case, then A would not have been the case' is true.

Now, the question is whether these two necessarily imply the following: 'If B and C were not the case, then A would not have been the case' is true. The truth conditions of the two assumptions according to Lewis's possible world semantics, which is what Cohen & Meskin use for truth conditions of counterfactual claim, are the following:

(1) Truth Condition of 1: the closest B-world is also an A-world.

(2) Truth Condition of 2: the closest C-world is also an A-world.

These two truth conditions do not imply that the closest B & C-world needs to be an A-world. The closest B & C-world could be farther away than both the closest B-world and the closest C-world, hence it may not be an A-world. Hence, the counterfactual account does not satisfy the Conjunction principle whereas the Dretskean account satisfies the Conjunction principle. In other words, in the counterfactual account a signal that carries the information that B and the information that C separately may not carry the information that B and C. This is counter-intuitive. Since the Conjunction Principle is intuitively correct and there

is no reason for rejecting or limiting its application, then not satisfying the Conjunction Principle is a deficiency of the counterfactual account. On the other hand, satisfying the Conjunction principle is a good feature of the Dretskean account.

In short, the counterfactual account is beneficial in terms of being in line with the limited application of the Xerox principle, but it leads to an unacceptable consequence with respect to the Conjunction Principle. On the other hand, the Dretskean account does not match the limited application of the Xerox principle, but satisfies the Conjunction Principle. There is no winner in this game; the result at best is a tie. Hence, it is not true that the information flow properties provide justification for preferring the counterfactual account over the Dretskean account as Cohen & Meskin claim.

5.2. Information and Laws. The second advantage of the counterfactual account, according to Cohen & Meskin, is that Drestke's account appeals to natural laws and hence requires a more expensive ontology. It is true that appeal to natural laws (or nomic dependencies between a signal and its informational content) is essential for the Dretskean account. The main rationale for this is to distinguish genuine information carrying relations from coincidental correlations. If your room and my room have the same temperature at a moment, the thermometers in both rooms will show the same temperature. Yet, it would be wrong to say that the thermometer in your room carries information about my room's temperature. For information carrying relations, there needs to be some lawful dependency between the number that the thermometer shows and the temperature of the room. This dependency holds between the thermometer in my room and my room's temperature. There is no such dependency between my room and the thermometer in your room. The nomic dependency requirement does not directly appear in Dretske's informational content definition. However, assigning unity to the conditional probability in the definition is a direct result of nomic dependencies. Dretske is very clear on this issue.

> In saying that the conditional probability (given r) of s's being F is 1, I mean to be saying that there is a nomic (lawful) regularity between these

> event types, a regularity which *nomically precludes* r's occurrence when s is not F. (emphasis original) [Dretske 1981, p.245]

On the grounds of ontological economy, an informational account that does not to appeal natural nomic dependencies is preferable to the ones which make such an appeal. Cohen & Meskin claim that their account does not have to make any such appeal. Counterfactuals, they say, are considered as presuming nomic dependencies between the constituents of counterfactuals [Goodman 1954]. But, they continue, this is not untendentious. That is to say, 'some think it is a mistake to characterize counterfactuals as essentially dependent on laws'. They do not make any commitment about the issue. They say that they are agnostic about whether counterfactual relations are essentially dependent on the existence of natural laws. This is why they claim that the counterfactual account is less ontologically committed than the Dretskean style probabilistic account. It is not clear at all whether their agnosticism is justified enough to give them the leeway of drawing such a conclusion. It is useful to visit their counterfactual claim again to make an assessment about this issue.

> **W:** x's being F carries information about y's being G if the counterfactual conditional 'if y were not G, then x would not have been F' is non-vacuously true.

In this definition, there is no direct reference to nomic dependencies. However, the definition is incomplete unless the truth condition of a counterfactual claim is specified, because in order to identify the existence of an instance of information carrying relation one needs to be able to assess the truth value of the relevant counterfactual claim. Specifying truth conditions of counterfactual claims means providing a semantics for counterfactuals. The commonly used semantics for counterfactuals is the possible worlds semantics. Once such a semantics is introduced, Cohen & Meskin's claim about less ontological commitment becomes controversial if not false. Which one is less ontologically committed: requiring natural laws or requiring possible worlds? The answer is not obvious at all.

In short, despite the fact that Cohen & Meskin's counterfactual definition does not make any direct reference to a specific ontology, their definition needs to specify truth conditions of counterfactual claims. Once this is done, the counterfactual claim becomes at least as

ontologically expensive as Dretske's probabilistic account. Cohen & Meskin may object to this conclusion by claiming that they want to be agnostic about the truth conditions of counterfactuals. I do not think that such a move is available for them since it would make their account incomplete. Moreover, even when such a move is granted to Cohen & Meskin, their account faces some other difficulties. When such a move granted, their counterfactual account becomes identical to Loewer's 1983 proposal about defining informational content with backtracking conditionals. That is to say, agnosticism move about truth conditions of counterfactuals would make Cohen & Meskin's account a variation of Loewer's proposal. Loewer proposed the following definition for informational content in his review of Dretske's account.

> r's being R carries the information that s is F iff r is R and if r is R then s must have been F.

The conditional claim in the definition is what Lewis (1979) calls 'a backtracking conditional'. There is no reference to laws or nomic dependencies in Loewer's proposal. Natural laws come into play only when Loewer defines truth conditions for backtracking conditionals. Loewer says,

> Truth conditions for these, as for other conditionals, are (approximately) 'there are laws L, conditions C which are cotenable with R(r) such that L&C&R(r) imply F(s).'

If Cohen & Meskin choose not to specify truth conditions for counterfactuals, then the same move should be available to Loewer as well. However, if this is the case Cohen & Meskin's counterfactual account becomes almost identical to Loewer's proposal with no significant difference since any counterfactual claim could be written as a backtracking conditional. Loewer's backtracking conditional 'if r is R then s must have been F' could be rephrased, without loss of meaning, as 'If s were F, then r would have been R'. This is nothing but Cohen & Meskin's counterfactual claim about information carrying relations. It should be noted that Cohen & Meskin discuss the similarity between their counterfactual account and Loewer's backtracking conditional based account. So, pointing out the similarity does not

add anything significant to the discussion. However, what is being claimed here is more than just pointing out the similarity. The claim is to show the consequences of choosing the agnosticism move with respect to truth conditions of counterfactuals. Moreover, I claim that specifying truth conditions is necessary for a complete counterfactual account of information carrying relations.

5.3. Doxastic States. In the Dretskean account, the information carried by a signal is relative to the background knowledge of the recipient of the signal. Such a reference comes naturally for Dretske for two main reasons. First, the reference fits to Shannon's analysis of information as uncertainty reduction. One's background knowledge surely determines the amount of uncertainty reduction that a signal provides to him. If you don't know that the city of Urfa is located in Turkey, then the signal 'Hilmi was born in Urfa' will not reduce your uncertainty about which country Hilmi was born. However, the same signal for another person who happened to know that Urfa is in Turkey will completely reduce the uncertainty about the country where Hilmi was born in. Not only does such a reference fit Shannon's framework, but it also matches our common intuitions about information flow.

Cohen & Meskin accept these features of referring to background knowledge in the definition of information carrying relations. However, they have another worry. The worry stems from the motive of having objective and naturalistic analyses of relevant philosophical concepts. The whole point of using information theoretic concepts is to provide a naturalistic account of mental content, belief and knowledge. If there is a reference to a semantic concept in the definitions of any of these concepts, then the result will be a circular and non-naturalistic account. Thus, the goal will not be achieved. Their worry is that referring to background knowledge makes Dretske's definition of informational content non-naturalistic and his definition of knowledge circular. Dretske has a quick and powerful answer to both of these objections. In each case, the reference to any semantic notion could be eliminated by backwards iteration. In other words, both of these definitions are recursive definitions. We have already seen Dretske's definition for informational content. The variable k in the definition shows the recursive character of the definition, and a backwards iteration will

provide the base of the recursion where there is no reference to background knowledge. The same reasoning applies to Dretske's definition of knowledge also. Dretske says that a continuing application of the analysis of knowledge and information will take us to the following point: 'we reach a point where the information carried does not depend on any prior knowledge about the source, and it is this fact that enables our equation to avoid circularity'. [Dretske 1981, p.86]

Cohen & Meskin do not find the backwards iteration reply to the circularity objection convincing. They see no reason to believe that Dretske's recursive definition has a base for all cases. Apparently, if that is the case, then Dretske's backwards iteration reply will not work. Given the structure of Cohen & Meskin's objection, the burden of proof lies on their side. They have to show at least one case where backwards iteration does not stop. They try to provide such an example by exploiting the possibility of two pieces of information mutually depending on each other. Here is their example,

> [L]et it be that, as Dretske claims, K's knowledge that s is F depends on the information that s is F and therefore (because of the role prior knowledge plays in his analysis of information) also on some other bit of knowledge K has about s (e.g., that s is G). For the same reasons, it seems entirely possible that K's knowledge that s is G depends on some further bit of knowledge K has about s. But nothing in Dretske's account rules out the possibility that this further bit of knowledge is in fact K's knowledge that s is F; on the assumption that the dependencies under discussion are transitive, an immediate regress ensues. [Cohen & Meskin 2006]

At first glance, the case of two mutually dependent two pieces of information is problematic for Dretske's theory. However, the situation becomes different when one asks under which conditions such a mutual dependence can occur. The mutual dependence can be a result of an analytic or a nomic connection between two pieces of information. In both of these cases, two pieces of information would be carried by the same signal in Dretske's theory. That

is to say, these pieces of information will be an instance of nested information. In other words, in such cases the mutual dependency that leads to regress does not exist, because these two pieces of information cannot be separated from each other. Dretske develops the notion of nested information exactly for such cases.

> For if a signal carries the information that s is F, and s's being F carries, in turn, the information that s is G (or t is H), then this same signal also carries the information that s is G (or t is H). For example, if r carries the information that s is a square, then it also carries the information that s is a rectangle.This point may be expressed by saying that if a signal carries the information that s is F, it also carries all the information *nested in* s's being F. (emphasis original) [Dretske 1981, 71]

In short, Cohen & Meskin's mutual dependency argument does not show that there are cases which will not lead to a stopping point in Dretske's backwards iteration that eliminates reference to doxastic states in his recursive definition of informational content.

5.4. Interim Conclusion. Cohen & Meskin claim that the counterfactual account that they propose should be preferred to the Dretskean probabilities account on three grounds: ontological economy, not referring to doxastic states and limited application of the Xerox principle. As I showed in the previous section, none of these is really an advantage. For the first two alleged advantages, Cohen & Meskin's objections to Dretske's theory are not well-justified. For the last one, their objection is justified, but their counterfactual account fails to satisfy another essential principle, i.e. the Conjunction Principle.

Given this analysis, there is no good reason for giving up Dretske's probabilistic account and adopting the counterfactual account. However, as the reader should remember, the main problem that led us to analyze the counterfactual account was another major problem that Dretske's theory had, i.e. the lack of a probability interpretation that grounds inverse conditional probabilities. Despite the fact that the alleged advantages of the counterfactual account are not real advantages, it is still the case that the probability interpretation is not a problem for the counterfactual account whereas it is a problem for the Dretskean account.

Hence, choosing the Dretskean account for defining informational carrying relations will not be well-justified until we offer a solution for the probability interpretation problem. The following section aims to achieve this goal.

5.5. Single Event Probabilities: The Solution. As previously discussed, the most plausible candidate for grounding the use of inverse conditional probabilities is the relative frequency approach where the probability of an event is defined as the limit of the sequence that this event belongs to. A direct result of this approach is that the probability concept is meaningful only in relation to sequences of events, not in relation to single events. This has been the main attack against the Dretskean way of defining informational content. The relative frequency approach is the probability interpretation that is commonly used in empirical sciences. Given this prevalent use, several attempts have been made for solving the single event probabilities problem. One of these attempts, the one developed by Reichenbach and defended by Salmon, is very promising and provides grounds for defending the Dretskean use of inverse conditional probabilities. If this is right, then the main argument that led to searching for a new definition for informational content on the basis of counterfactuals or backtracking conditionals will lose its significance.

Reichenbach's book, *The Theory of Probability*, presents an elegant defense of the relative frequency approach [Reichenbach 1949]. In his analysis of single events, he claims that the concept of probability is fictitiously applied to single events. That is to say, there is no proper probability attached to single events. However, whenever a single event is presented, the probability concept is extended to such an event by finding the probability associated with a 'proper' infinite sequence and transferring that value to the given single event. The problem here is to specify what is meant by 'proper', because a single event belongs to many different sequences and the probabilities associated with these sequences may differ significantly. Let's state this formally.

In the relative frequency approach, probability is defined between two classes: Pr (A , B). B refers to the attribute class, that is to say the event that we want to assign a probability value. A refers to the reference class which represents the sequence that the

attribute event belongs to. In the case of single events, we have a clear description of the attribute class: the event of getting an ace from a deck of cards, the event of having a woman president of the United States of America in 2008, what premium a given individual should be charged for his car insurance etc. The problem is to identify the reference class. This is a practical problem and many 'pragmatic' solutions are being used by several institutions, for example insurance companies. They try to assign a category to a given individual, and then assess the probability of having a car accident for that category. This probability is accepted as the given individual's probability of being involved in an accident. To find the proper category requires identifying relevant properties. Whether the individual is a male or female or the person age does matter for insurance purposes, but the day of the week that the person was born on hardly matters. Reichenbach, for finding the proper class, proposes that we should choose 'the narrowest reference class for which reliable statistics are available.' Wesley Salmon suggests that one needs to choose the broadest homogenous reference class of which the single event is a member. Salmon's argument for his claim provides a good justification. Since any kind of reference class attribution is necessarily an inductive claim, it is better to broaden the basis of our inductive inference. Moreover, Salmon claims, the reference class should not be broader than it needs to be. That is to say, we want to include relevant classes not irrelevant ones. He formalizes his notion of statistical relevance in the following way.

> Suppose we ask for the probability that a given individual x has a characteristic B. We know that x belongs to a reference class A in which the limit of the relative frequency of B is p. If we can find a property C in terms of which the reference class A can be split into two parts $A \bigcap C$ and $A \bigcap \sim C$ [where $\sim C$ is the complement set of C], such that $P(A \bigcap C) \neq P(A, B)$. Then C is statistically relevant to the occurrence of B within A. (Salmon p.91)

This formal definition provides statistically relevant classes. If there is no relevant partition C, then A is a homogenous class that we need for identifying the probability of the event B.

For example, the probability of getting heads is 1/2 with respect to tosses of a coin. There is no relevant subdivision of the reference class of tosses of a coin. For example, when one considers the tosses that are made at night versus at morning hours, still the probability will be 1/2. Hence, $P(A \bigcap C, B) = P(A, B)$ where C stands for the time of the day of coin tosses. Since there is no relevant subdivision, we can safely conclude that the reference class is a homogenous one, and it is the broadest among all other homogenous reference classes.

Salmon's solution for assigning probability values to single events seems to be a better candidate than Reichenbach's. However, it should be noted that which one of these suggestions, i.e. the narrowest or the broadest, is more accurate is not very crucial for my position here. The main point is that there is a way to assign probabilities to single events even if it is not probability proper. This could be done either in Reichenbach's way or Salmon's way.

6. Conclusion

In conclusion, the use of inverse conditional probabilities, i.e. the first person perspective, brings advantages, which would not be available otherwise, for constructing a theory of mental content. The single event probabilities problem could be accommodated by following Reichanbach's and Salmon's footsteps. On the other hand, the counterfactual suggestion for defining informational content does not bring any serious advantage to the table. Moreover, it brings a huge metaphysical baggage because of the very notion of counterfactuals. Thus, I incorporate the use of inverse conditional probabilities in the theory that I develop. As I have mentioned, Dretske's 1981 framework is the first attempt at using inverse conditional probabilities in the mental content literature. However, his insistence on assigning unity to conditional probabilities reverses the first person perspective to the worst version of the third person perspective, i.e. the omniscient third person perspective. His insistence might seem unjustified at first glance. However, he has powerful reasons for doing that. In the following chapter, I analyze his reasons and motivations for doing that, and provide arguments against his reasons. Besides these, I also provide a positive definition for informational content that uses inverse conditional probabilities without assigning unity to conditional probabilities.

CHAPTER 4

First Step Towards a Probabilistic Theory of Mental Content

In his book, *Knowledge and the Flow of Information*, Dretske develops a theory of mental content based on the notion of informational content. By using this notion, together with the tools of the mathematical theory of communication developed by Shannon and Weaver, Dretske aims to give an account of mental content, perception, belief and knowledge. There are several philosophically valuable claims, mistakes and problems in Dretske's theory, and since the publication of his book a sizeable literature has been produced on these issues. This chapter doesn't aim to give a full analysis of Dretske's theory; rather it aims to discuss the fundamental notion of his theory, informational content, and problems associated with it. After discussing these problems, I provide the probabilistic definition that I use for the theory that I develop in this dissertation.

Dretske defines the notion of informational content, which is the basis of his claims regarding mental content, perception, belief and knowledge, as following.

> **Informational Content:** A signal r carries the information that s is F = the conditional probability of s's being F, given r (and k), is 1 (but, given k alone, less than 1) [k refers to background knowledge] [Dretske 1981, p.65]

Assigning unity to the conditional probability in the definition has severe implications. For example, as a result of the unity, Dretske rejects the possibility of partial information and misinformation. He says that *'information is certain; if not, it is not information at all'*. This result sets the stage for a problem which makes Dretske's theory unsuccessful: the problem of misrepresentation. A successful theory of mental content must make room for the cases where misrepresentation occurs. Unfortunately, Dretske's theory is not able to do so, and this inability has its roots in the denial of misinformation and partial information.

Because of these unacceptable consequences, many scholars have questioned the legitimacy of assigning unity to conditional probabilities. The 1983 BBS article and the 1987 special issue of *Synthese* on Dretske's information theoretic approach give us a valuable representative collection of these criticisms. In his response to these criticisms, Dretske says that despite these immense criticisms, no one was able, nor did they even try, to reject his arguments for assigning unity to conditional probabilities [Dretske 1983, pp.84-85]. What Dretske says is true; I could not find any article which attempts to reject Dretske's three arguments - the Xerox Principle, the Arbitrary Threshold Thesis and the Conjunction Principle - for his controversial claim regarding conditional probabilities.

In their commentaries in the BBS article mentioned above, both Lehrer and Kyburg clearly stated the implausibility of Dretske's arguments, but to my knowledge they have not attempted to give a full account of why these arguments need to be refuted. Without such a rejection, Dretske's claim regarding conditional probabilities still stands. Needless to say, it stands with its unacceptable consequences.

Another attempt that needs to be mentioned here is an unpublished manuscript by Scarantino [2005] of the University of Pittsburgh. Cohen & Meskin [2006] claim that Scarantino refutes Dretske's arguments for supporting his claim regarding conditional probabilities. However, in his manuscript, Scarantino tries to refute only one of Dretske's three arguments: the one based on the Xerox principle. He tries to do so by using Fano's [Fano 1961] measure for the amount of information that a signal carries instead of the one that Dretske uses for his theory. The details are not important here. Suffice it to say that Fano's measure is one of the mathematically legitimate measures of the amount of information (entropy in technical terminology), but it has nothing to do with rejecting Dretske's Xerox principle. Hence, Scarantino's manuscript represents a wrong step in the right direction. Moreover, his manuscript does not even mention the other two arguments, i.e. the Arbitrary Threshold Thesis and the Conjunction Principle.

Besides rejecting Dretske's arguments for assigning unity to conditional probabilities, a positive suggestion for how to revise the definition of informational content is also needed

if one wants to exploit the notion of information and the mathematical theory of communication for philosophical problems associated with mental content. The situation in the literature in regards to making a positive suggestion is a little bit better. Usher [2001] introduced a statistical theory of mental representation by rejecting the necessity of assigning unity to conditional probabilities. However, he did not locate his suggestion in the context of Dretske's original arguments. Hence, it is difficult to assess the success of his suggestion. Moreover, his suggestion is not able to capture some of Dretske's original motivations which are philosophically valuable. In a similar manner, Eliasmith in his unpublished dissertation introduced a theory of mental/neural content where the conditional probability of a mental entity given its referent is less than one. He also developed his theory without rejecting Dretske's original arguments. Moreover, his theory is essentially a two-factor theory, so it carries all the problems that come with a two factor approach.[1]

The overarching goal of this chapter is to achieve the two needs mentioned above: to reject Dretske's three arguments for assigning unity to conditional probabilities and to construct a better definition for the notion of informational content. The latter will serve as the basis for the probabilistic theory of mental content that I develop in this dissertation (Please see Chapter 6). On the way to achieving these two goals, I also analyze one theoretical motivation, which is not directly stated in Dretske's works but hidden between the lines, that led him to assign unity to conditional probabilities. This motivation is the distinction that Dretske borrows from Paul Grice - the distinction between natural meaning versus non-natural meaning. Not only do I identify and analyze this underlying motivation, but I also criticize Dretske's interpretation of the distinction and offer a better one.

Given these goals, this chapter is organized as follows. In Section 1, I state Dretske's arguments and reject them one by one. The Gricean distinction between natural meaning and non-natural meaning that serves as the main theoretical motivation for Dretske is analyzed in Section 2. The third section provides a positive suggestion for the notion of

[1]The reasons for rejecting two factor approaches, and the necessity of adopting a causal/informational framework are elaborated in the first chapter of this dissertation.

informational content. The notion is defined in terms of ordinal rankings of the conditional probabilities all relevant alternatives.

1. Arguments that Survived 25 Years

In the definition of informational content, the conditional probability of the signified (s's being F) given the sign (together with the required background knowledge) is 1. As mentioned above, this claim received many criticisms, but no one has attempted to refute the arguments that Dretske offers for supporting it. In this section, I aim to refute Dretske's original arguments, and show that Dretske's original expectations from a definition of the notion of informational content can be fulfilled by assigning something less than one to conditional probabilities.

In his *Knowledge and the Flow of Information*, Dretske presented three arguments for claiming that the value of conditional probability in his definition of informational content must be one - nothing less. The first one is about the transitivity of information. He claims that information flow is possible only if the flow is transitive, i.e. if a signal A carries the information B, and if B carries the information C, then A must carry the information that C. It is a simple mathematical fact that conditional probabilities are not transitive unless they are equal to one (please see Appendix 1). Hence, in order to satisfy the transitivity property, i.e. the Xerox Principle, conditional probabilities must be one.

Secondly, Dretske says that *'there is no arbitrary place to put the threshold that will retain the intimate tie we all intuitively feel between knowledge and information.'* If the information that 's's being F' can be acquired from a signal which makes the conditional probability of this situation happening something less than 1, say for example 0.95, then 'information loses its cognitive punch.'[Dretske 1981, p.63]

The principle that he uses for his third argument is a close relative of the Xerox Principle, and he calls it the Conjunction Principle. If a signal carries the information that B with a probability of p_1 and the information that C with a probability of p_2, the probability of carrying the information that B and C must not be less than the lowest of p_1 and p_2. [2]

[2] p_1 is $P(B|S)$ which stands for the probability of B given the signal. Likewise, p_2 is $P(C|S)$.

However, again it is a simple mathematical fact that this could not happen with conditional probabilities if they are less than one (please see Appendix 2).

As a result of these three arguments, Dretske claims that the conditional probability in his definition of informational content must be one. Before analyzing these arguments one by one, let me briefly mention Dretske's general motivation for offering these arguments. To put it very simply, Dretske thinks that the connection between the world and the mind should be explained by using the notion of information because, after all, the main function of our perceptual mechanisms is to learn about our surroundings - the crucial concept being 'to learn'. Learning means acquiring information. Hence, the notion of information is the key for demystifying our perceptual connection with the world. Dretske's arguments become more intuitive once thought of as a result of learning metaphor. If I can learn B from A and C from B then I should be able to learn C from A. This intuitive claim is nothing but the Xerox Principle. 'Learning B from A' is identical to 'A carries the information that B'. Likewise, 'Learning C from B' means 'B carries the information that C'. These two together imply that 'I can learn C from A', i.e. A carries the information that C. A similar reasoning applies to the Conjunction Principle. For the Arbitrary Threshold Thesis, since the metaphor is to learn, ideally we want to learn the truth not an approximation of truth. In short, Dretske's intuitive motivation for his arguments is the metaphor of learning.

Now, let me analyze his arguments one by one.

1.1. The Xerox Principle. As mentioned above, Dretske's Xerox principle [Dretske 1981, pp.57-58] relies on a claim regarding information flow. He says that it is transitive. This property is, in Dretske's own words, absolutely fundamental for information flow. However, conditional probabilities do not obey the transitivity property unless they are equal to one. For example, when the probability of B given A and the probability of C given B both are above 0.9, the probability of C given A could be way below 0.9; it could even be zero. So, Dretske concludes that conditional probabilities must be one. This is what leads him to deny the possibility of partial information and misinformation. Once

again, his motto is *'information is certain; if not, it is not information at all.'* Dretske's argument for assigning unity to conditional probabilities simply goes as following.

Premise 1 The Xerox principle is fundamental.

Premise 2 The Xerox principle implies that conditional probabilities must be one.

Therefore, conditional probabilities must be one.

He does not offer any argumentation for his first premise. He simply assumes it, but for the second premise he attempts to give a reductio proof. As his reductio assumption, he supposes that a signal can carry information in the face of positive equivocation, i.e. the required conditional probability is less than one (please see the definition of informational content on p.76). Then in his rather lengthy proof, given the reductio assumption, he derives the following: in an informational chain, A could carry the information that B, and B could carry the information that C where A may not carry the information that C, he says, is absurd because it contradicts the Xerox Principle. His conclusion relies on the Xerox principle. In other words, the rationale that he offers for the second premise is nothing but the first premise of his argument. If there were no reason for doubting the validity of the Xerox principle, then his argument would be acceptable. However, there are good reasons for questioning the validity of Dretske's fundamental principle.

The Xerox principle, obviously, works in idealized situations. In an artificial system of coding, the principle will hold. Such examples could easily be found in the stock of military communication systems where both the source and the receiver have perfect knowledge of the coding system with almost noiseless channels. However, this idealization is hardly ever true in real life applications of information flow. Dretske himself mentions such an example in a footnote. The footnote which is intended to support his reductio proof mentioned in the previous paragraph turns out to be an unintended confession. The example in the footnote has three information sources (A, B and C) and eight events for each source ($\{a_1, \ldots, a_8\}$ at A, $\{b_1, \ldots, b_8\}$ at B and $\{c_1, \ldots, c_8\}$ at C) with specific probabilities. There is an informational link between these three different sources; as a result, the events at each source are linked to the events of other sources with some conditional probabilities. The conditional probabilities are set up in a way that there is some equivocation at each source,

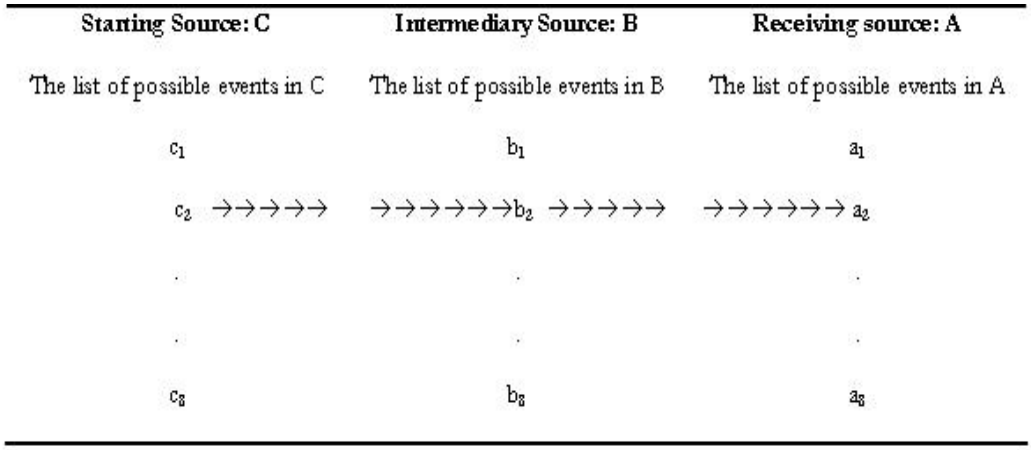

FIGURE 1. Dretske's Footnote

i.e. conditional probabilities are less than one. Then he takes one event at C, c_2, and follows the flow of the information that c_2 occurred at C to B with the event b_2 and then to A with the event a_2. Figure 1 shows the flow. Dretske calculates the amount of information that flows through the channel in the face of equivocation. Without bothering the reader with the calculation and formula details (for these details please see Appendix 3), let me quote Dretske's final claim.

> c_2 occurs at C, generating 3 bits of information ... b_2 carries approximately 2.78 bits of information about what occurred at C. a_2 carries 2.78 bits of information about what occurred at B. If, however, we calculate the equivocation between A and C, we find that it is 1.3 bits. a_2 carries 1.7 bits of information about what happened at C. [Dretske 1981, p.254]

As obvious from this quote, there is information flow from C to A through B despite the fact that the required conditional probabilities are less than one. It is true that there is some information loss, because 3 bits of information at C becomes 1.7 bits at A. However, there still is flow. This example, as opposed to Dretske's original intention, clearly shows the possibility of information flow without conditional probabilities being one or in other words without satisfying the Xerox principle.

The data processing inequality theorem of the mathematical theory of communication is directly related to Dretske's Xerox principle, and it shows that the Xerox principle holds only in idealized situations.

> **Data Processing Inequality Theorem:** If there is an information flow from X to Z through Y, then the mutual information between X and Y is greater than or equal to the mutual information between X and Z. More formally:
> $If X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$ [Cover 1991, p.32][3]

The equality condition in the greater than or equal to relation between I (X ; Y) and I (X ; Z) holds only if the chain formed by the information from X to Z through Y ($X \rightarrow Y \rightarrow Z$) is a Markov chain. A Markov chain occurs when the conditional distribution of Z depends only on Y and it is independent of X. Obviously this is a very strict constraint, and rarely true in real life information channels. If this constraint is not fulfilled, then the probability of having equality becomes lower and lower as the chain of information flow becomes longer, because of having more constraints. Hence, Dretske's conclusion about conditional probabilities is valid only in idealized cases. It should be noted that Dretske's strict Xerox principle corresponds to the equality between I (X ; Y) and I (X ; Z) in the data processing equality theorem, not the greater than relation.

Markov chains, i.e. informational chains where only the two subsequent members of the chain conditionally depend on each other, are not strong enough to exploit the statistical regularities that may exist in the informational source. Shannon in his seminal article, *The Mathematical Theory of Communication*, showed the importance of longer conditional dependencies in a sequence for exploiting the statistical regularities in an informational source [Shannon & Weaver 1980, pp.43-45]. The informational source that he chose was English. As it is known, some letters are more frequent than others in English. This is

[3]The theorem is rephrased for the sake of simplicity. The notion of mutual information has not been discussed yet. It is a crucial notion for my dissertation project, but for the purposes of this chapter suffice it to know that it is the amount of information that two events carry about each other. More technically, it is a similarity measure between two random variables. This notion will be discussed in the following chapters in detail.

the main reason for assigning the highest point value to the letter Q in Scrabble; it is the least frequent letter in English words. This is an important statistical regularity of English, but not the only one. There are also patterns depending on the previous letters that occur in a sequence. For example, the probability of having an 'S' after an 'I' is different from the probability of having a 'C' after an 'I'. Similarly, the probability of having a 'U' after the sequence 'YO' is different than the probability of having an 'R'. Shannon used these statistical patterns in sequence in order to produce intelligible sequences in English without feeding any extra rule to the sequence producing mechanism. For all sequences, he assumed a 27-symbol alphabet, the 26 letters and a space. In the first sequence, he used only the occurrence frequencies of letters; he called this first order-approximation. The idea behind the process by which he produced the sequence can be thought in the following way. Imagine a 27 sided die where each side is biased according to its occurrence frequency. Then, by simply rolling the die at each step one decides the symbol that should appear for that step. The output of his first sequence where only letter frequencies used is the following.

First Order Approximation

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

For the second sequence, the frequencies that he used were the frequency of a letter given the letter that comes right before E. That is to say, instead of using the simple occurrence frequency of the letter E, he used the conditional frequency of E given the previous letter, for example if the previous letter is K, then he used the occurrence frequency of E given K. This is his second-order approximation.

Second Order Approximation

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

In the third order approximation, he used the occurrence frequencies of letters given previous two letters instead of one.

Third Order Approximation

	Meaningful Sequences (MS)	The length of MS	The total length	The Success Index	% Increase
1st order	–	0	72	0	NA
2nd order	ON, ARE, BE, AT, ANDY	13	118	0.11	NA
3rd order	IN, NO, IN NO, WHEY, OF, OF, THE, OF THE, IS, OF	30	108	0.28	%154

FIGURE 2. Improvement Index

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

There is an improvement from the second-order approximation to the third-order. This improvement may not seem significant at first glance. However, when measured quantitatively the third-order approximation almost triples the success of the second-order approximation, and quantitative measures are the proper way of identifying such improvements. Unfortunately, Shannon did not provide such a quantitative success index, because for his purposes the improvement was noticeable enough. A simple success index that can be used is the ratio of the length of the meaningful sequence to the length of the entire sequence. The success index values calculated accordingly are shown in Figure 2. As the table shows, the index value of the third-order approximation is equal to two and a half times of the index value of the second-order approximation. That is to say, there is a significant improvement from the second to the third order, and the level of improvement increases exponentially when one moves to higher order approximations such as the fourth-order, the fifth-order and so on. In short, the sequence of English letters becomes much more meaningful when one increases the length of the dependencies in conditional probabilities. In other words, a successful use of statistical regularities requires longer informational chains where the conditional probability of an entity depends not just on the previously occurring one, but on several others that come before that entity. Shannon's second order approximation, conditional probabilities given just the previous letter, corresponds to the idea of Markov chains mentioned above. Dretske's strict Xerox principle presumes a Markov chain and hence stops at the second order approximation level. However, the amount of information that one can

exploit from an informational source (in the case of mental content, the source is the world) by a Markov chain, is very limited, as shown in Shannon's second order approximation. Most of the informational sources (for example, natural languages and the external world) are much richer, and to exploit such richness one needs to extend dependencies beyond the limits of a Markov chain. As Figure 2 shows, even going one order level up from a strict Markov chain significantly increases the ability to exploit regularities in an informational source. Hence, a more lenient version of Dretske's Xerox Principle is required.

Equating conditional probabilities to one has another negative effect for Dretske's theory. One of Dretske's reasons for using the mathematical theory of communication, and thereby the notion of information, is to make use of the statistical properties in the external world as the main source for mental content. If, however, conditional probabilities are one, then the theory becomes a strictly logical theory, and loses its statistical character. In a sense, by equating conditional probabilities to one, Dretske is sawing off the branch he is sitting on.

There are two ways to see the logical character of Dretske's theory. The first one is about logical implication. In Dretske's approach, a mental representation, say R, indicates (or carries information) that S -an external state of affairs- with a probability of one. This, together with the existence of R in my mental realm, implies the existence of S in the external world. In other words, the following statement is true, and it makes Dretske's theory lose its statistical character and become closer to a logical structure.

$(Pr(S|R) = 1$ and $R) \rightarrow S$ ($\rightarrow$ symbolizes logical implication)

The second way of seeing the logical character is the similarity between the Xerox principle and a well-known rule of natural deduction: the hypothetical syllogism. The similarity between the following formulation of the Xerox principle and the hypothetical syllogism is obvious.

Xerox Principle

A has the information that B.

B has the information that C.

Therefore, A has the information that C.

Hypothetical Syllogism

A implies B.

B implies C.

Therefore, A implies C.

A further implication of collapsing the information theoretic approach to a logical structure is that it makes generation of new information impossible. As Bar-Hillel points out [Bar-Hillel 1955], logical truths do not generate any information . Any logical truth is vacuously true, and any valid argument is equivalent to a logical truth constructed out of its premises and conclusion by using conjunction and implication. Hence, if we want to preserve the possibility of generating extra information (and thereby new knowledge), we should question the logical character of Dretske's theory, and hence question assigning unity to conditional probabilities.

The moral of the story is that the Xerox Principle, as it is stated by Dretske, is not well justified. It works only for a very limited range of cases. On the other hand, as both the *Data Reduction Theorem* and Dretske's unintended confession suggest, it is possible to have information flow without assigning unity to the conditional probability of the signified given a sign. Moreover, this possibility provides more power for exploiting statistical regularities in an information source and more importantly makes it possible to generate new knowledge.

1.2. The Arbitrary Threshold Thesis. As mentioned above, Dretske's overall aim for using the notion of informational content is to be able to define mental content, beliefs and thereby knowledge in naturalistic terms. Since one of his aims is to achieve a naturalistic explanation for the necessary and sufficient conditions of knowing something, and in traditional epistemology knowing something requires that something to be true, he tries to provide an intimate connection between the notion of information and truth. On the other hand, since Kyburg introduced the lottery paradox [4] in 1961, it is commonly accepted that

[4]Briefly the lottery paradox is the following. If any probabilistic evidence were enough for knowledge, then the evidence for not winning in a lottery would be enough. Assume that one person bought a ticket for a lottery of 1 million tickets. The probability of not winning is very high. Moreover, assume that this probability is enough for knowledge. Then he 'knows' that he is not going to win, because the likelihood

no arbitrary threshold of probabilities is sufficient for claiming that knowledge based on a probabilistic account implies truth [Kyburg 1961]. Even the highest probability that can be imagined (other than unity) does not entail the true statement that is required for knowing that the event will occur.

This fact about the lack of a probabilistic threshold for a proper knowledge claim can be proven formally. In fact, Lehrer, in his 1970 article *Justification, Explanation and Induction*, provides such a proof. In this article, Lehrer constructs a formal inductive system that is based on the notion of conditional probabilities. This makes Lehrer's notion of inductive inference a close relative of Dretske's notion of informational content. Because of this close affinity, it is fruitful for the purposes of this project to briefly visit Lehrer's formal system [Lehrer 1970].[5]

Lehrer starts with the notion of justification which is one of the three main criteria (justification, truth and belief) for knowledge in traditional epistemology. A proper justification requires a good explanation. A good explanation must obey three main principles, according to Lehrer. These three principles are the following: i) the Conservation Principle: if a hypothesis is explained by our background knowledge, then this same hypothesis is also explained by any extension of our background knowledge. ii) the Conjunction Principle: this principle simply says that if two hypotheses are explained by our background knowledge separately, then so is the conjunction of these two hypotheses.[6]iii) the Consistency Principle: the set of hypotheses that are explained by our background knowledge must be logically consistent.

---

of winning is very small. This reasoning is true for every person that has bought a lottery ticket. Thus, everyone participated in the lottery 'knows' that they will not win. On the other hand, at least one person has to win the lottery. Hence, the paradox ... The lottery paradox is formally discussed in the last section of this chapter

[5]The importance of Lehrer's framework and its fundamental notion become more apparent in Section 3 where I develop a better definition for informational content

[6]Lehrer has two different explanation principles from which he derives the conjunction principle. He needs to do that for technical reasons. For the sake of simplicity, I jump to his concluding principle, i.e. the conjunction principle, instead of starting with his two original principles.

In order to reach a formal system, Lehrer needs to specify the notion of explanation with more concrete notions. For this, he introduces another level, the inductive inference level. This level is the basis for the explanation level. In other words, the explanation level is grounded on the inductive inference level.[7] For a hypothesis to be explained by background knowledge, this hypothesis must be inductively inferred from the background. In turn, he tries to define the notion of inductive inference by using conditional probabilities. For example, a hypothesis, say h, can be inductively inferred from a set of background knowledge, say b, if the conditional probability of h given b - $P(h|b)$ - is above some specific value. In short, he has three levels: explanation, induction and probabilities. The most abstract one is the explanation level and the most concrete one is the probabilities level.

The connection between the explanation level and the inductive inference level makes it possible to reformulate three explanation principles in terms of inductive inference. For example, take the conjunction principle. If two hypotheses are inductively inferred from background knowledge separately, then the conjunction of these two hypotheses could also be inductively inferred from the background knowledge.

A proper definition of inductive inference by using conditional probabilities must obey the basic rules of probability and it should satisfy the main three explanation (inductive inference) principles mentioned above. Before giving the definition that he defends for inductive inference, Lehrer shows that any definition based on an arbitrary threshold does not work. He takes the following formulation for the inductive inference in order to prove this claim.

I (h,b) if and only if P (h | b) is at least m/n (Where the locutions 'I(h,b)'

[7]Lehrer focuses on inductive explanation because he thinks that any deductive explanation requires the explanandum as one of its premises, hence it is circular. Whether he is right or not about deductive explanation is not relevant for my current purposes. There is only one thing that I want to mention. His argument for showing that any deductive explanation is circular relies on a logical equivalence claim. He says that E is equivalent to the conjunction of F and (F → E). This is simply not true. Anyway, none of the deductive explanation issues is relevant for my current purposes.

‘ $P(h|b)$’ mean ‘h may be inductively inferred from b’ and ‘the probability of h given b’ respectively and m is a positive integer that is less than n which is another positive integer.)

The quantity m divided by n gives us a number less than one and greater than zero, and it serves as the threshold for the legitimacy of an inductive inference. Since no other specification is made for m and n, it serves as an arbitrary threshold. Lehrer's proof shows that no matter what m/n is, this definition leads to a contradiction. More specifically, it leads to allowing us to infer an impossible event from a set of background knowledge. His proof is just a formalization of the lottery paradox.

Despite the fact that Lehrer's proof supports Dretske's claim about the arbitrary threshold, this does not mean that conditional probabilities must be assigned to unity in the definition of informational content (or in the definition of inductive inference in Lehrer's system). There is another possibility. Instead of using a threshold, one could use an ordinal ranking of all possible conditional probabilities and pick the highest one for assigning an informational content to a signal (or for a proper inductive inference). Lehrer follows such a strategy, and defines inductive inference by using ordinal rankings of conditional probabilities. The definition that he uses is a little bit technical, but it simply comes down to the following idea. A hypothesis, h, can be inferred from a specific background, b, given that $P(h|b)$ is greater than $P(k|b)$ where k is any competing hypothesis.[8] Lehrer successfully proves that such a definition for inductive inference satisfies both the three main principles of explanation (the Conservation, the Conjunction and the Consistency principles) and the basic rules of probability calculus. Lehrer's success, at the least, is an existence proof for showing that an ordinal ranking of conditional probabilities can be used in the definition of informational content. That is to say, what Dretske infers from the no arbitrary threshold claim is not correct. **It is true that there is no arbitrary threshold for assigning a proper content to a signal, but this does not necessarily imply that conditional probabilities must be assigned to unity.**

[8]Obviously, being able to list all competing hypotheses is a challenge. Lehrer meets this challenge with the notion of minimally inconsistent set. His solution works despite the fact there is a minor problem in his definition. Section 3 explains this minor problem and the notion of minimal inconsistency.

Using the highest one among all possible conditional probabilities (conditional probabilities of possible contents, given a particular signal) provides a satisfactory framework, but it comes at a price. The price is to lose the possibility of absolute knowledge, i.e. the possibility of the traditional claim that knowledge implies truth[9]. The reason for that is the following. In the ordinal ranking approach, we will be selecting the hypothesis with the highest probability which, in most cases, is less than one. This is an acceptable consequence. The traditional project of epistemology requires absolute knowledge that could serve as the foundation of a network of knowledge claims. So, it is only normal to have absolute knowledge only for a very limited case. As mentioned above, unity for conditional probabilities happens only when a Markov chain is present. For all other cases, following Lehrer's suggestion, we have knowledge that is less than perfect. One may object to this by referring to cases like the lottery paradox that result from less than perfect knowledge. However, as Kyburg states, the point of the lottery paradox is not to establish a need for absolute knowledge. He says that the point is to show that there is no place for absolute knowledge in the body of empirical knowledge. And absolute knowledge, in his words, is 'as uninteresting as the knowledge of Absolute'. Moreover, he adds, most epistemologists 'took the lottery paradox to be a paradox to be explained away in conventional terms, rather than an indication that there might be something wrong with conventional epistemology' [Kyburg 1983, p.78]. The mistaken part of conventional epistemology is its insistence on absolute knowledge, which can be attained only in a few idealized cases.

Before concluding the discussion about Dretske's arbitrary threshold argument, it is relevant to mention that even a probability of one is not sufficient for entailing truth. Carnap & Jeffrey [1971] in their book, *Studies in Inductive Logic and Probability*, showed that a probability of one entails truth only for finite languages (or systems). So, even if we accept Dretske's insistence on the conditional probability of 1, he still needs to show that the system of mental states or the system of all possible informational contents is finite. It is not wrong to claim that an effort of proving such a thing would be absolutely futile.

[9]The other option is to adapt a weaker theory of knowledge

1.3. The Conjunction Principle. Dretske's third argument for assigning unity to conditional probabilities is the Conjunction Principle [Dretske 1981, p.100-101]. He says that 'if K knows that P and knows that Q, then he knows that P and Q (call this the Conjunction Principle)'. Since he tries to give a naturalistic account of knowledge by using informational content, then this notion should also obey the conjunction principle. That is to say, if a signal carries the information that B and if it carries the information that C, then it has to carry the information that B and C. However, when conditional probabilities are applied, this principle is violated except in the cases of the probability of 1 (please see Appendix 2).

The ordinal ranking approach, however, can accommodate this principle without assigning unity to conditional probabilities. Following Lehrer's suggestion we could use the following formulation: a signal, say r, has the content h_1 if and only if the conditional probability of h_1 given r is greater than the conditional probability of any other competing (all other h_is) hypothesis given r . This gives us the following inequality.

Ineq1 $P(h_1|r) > P(h_i|r)$ for all $i \neq 1$

Now, imagine that the signal has some other content as well, say k_1. The above definition gives us another inequality.

Ineq2 $P(k_1|r) > P(k_j|r)$ for all $j \neq 1$

Since the signal r carries the information that h_1 and it also carries the information that k_1, it should also carry the information that h_1 and k_1, according to the Conjunction Principle. In this ordinal ranking definition, this principle is easily met. In the new conjunction case, one needs to define the set of competing hypotheses. When only h_1 is in question, the competing hypotheses are $\{h_2, h_3, h_4, \ldots, h_n\}$. For the second case, the set is $\{k_2, k_3, k_4, \ldots, k_m\}$. For the conjunction case, since h_1 competes with h_is and k_1 competes with k_js, naturally the set of competing hypotheses is $\{h_2\&k_2, h_2\&k_3, h_3\&k_2, h_3\&k_3, \ldots, h_n\&k_m\}$. Once the set of competing hypotheses is defined, then to test whether the ordinal ranking approach satisfies the Conjunction Principle requires finding out if the following inequality is implied by Ineq1 and Ineq2.

Ineq3 $P(h_1\&k_1|r) > P(h_i\&k_j|r)$ for all $i, j \neq 1$

Without the notion of minimal inconsistency, Ineq1 and Ineq2 do not imply Ineq3, because conditional probabilities are not additive. The lack of this property causes the same problem explained in Appendix 2. However, the notion of minimal consistency imposes an independency constraint between h_1 and k_1. The hypothesis k_1 cannot be a competing hypothesis with h_1 and vice versa. For the purposes of a reductio proof, let's assume that k_1 is a competing hypothesis for h_1. If that were the case then $Pr(h_1|r)$ would be greater than $Pr(k_1|r)$ according to Ineq1. However, since h_1 and k_1 are competing with each other, h_1 would be a member of the competing hypotheses of k_1 as well, and then according to Ineq2, $Pr(k_1|r)$ would have to be greater than $Pr(h_1|r)$. That is to say, k_1's being a competing hypothesis for h_1 implies the following two claims:

$Pr(h_1|r) > Pr(k_1|r)$ and $Pr(k_1|r) > Pr(h_1|r)$

This is a plain contradiction. That is to say, the reductio assumption is wrong. Hence, k_1 cannot be a member of the set of competing hypotheses for h_1. Similarly, h_1 cannot be a competing hypothesis for k_1, either. The reductio proof applies to this case as well without the loss of generality.

The independence constraint between k_1 and h_1 as being two pieces of information carried by r makes Ineq1 and Ineq2 additive against the new set of competing hypotheses, i.e. $\{h_2\&k_2, h_2\&k_3, h_3\&k_2, h_3\&k_3, \ldots, h_n\&k_m\}$. Hence, Ineq1 and Ineq2, together with the notion of minimal inconsistency, imply the required inequality Ineq3. This suffices to show that the Conjunction Principle can be satisfied in a system where conditional probabilities are less than one. Please see Appendix 4 for a diagrammatic explanation of the independence constraint and its result.

Lehrer also provides an existence proof for meeting the challenge of the conjunction principle in his system based on the notion of minimal inconsistency. Although the explanation for the conjunction principle that I just stated has similarities with his proof, my explanation is different from his. In this sense, the explanation given above is original. Needless to say, Lehrer's proof provided fundamental insights for my explanation.

2. The Hidden Motivation

Besides his three arguments discussed in the previous section, Dretske also has a motivation for assigning unity to conditional probabilities. Dretske apparently wants to analyze mental content within the framework of Paul Grice's distinction between natural and non-natural meaning. He considers mental content and contents of propositional attitudes as similar to non-natural meaning instances, and he tries to naturalize contents of mental states by reducing them to natural meaning instances via the notion of informational content. Thus, it is necessary to analyze the Gricean distinction in order to fully understand both merits and shortcomings of Dretske's theory.

The idea of natural meaning arises from natural signs. Natural signs have their meaning without any assistance from human beings. For example, the existence of an oasis means that there is water around, the direction of a shadow means that the sun is in the other direction and so on. The main motivation of introducing natural meaning is to differentiate naturally occurring signs, which do not require any assistance from us, from non-natural signs that are formed through some sort of conventions. Any natural language or any code of human communication is a good example of non-natural signs and non-natural meaning.

In Grice's original distinction, natural meaning instances imply the truth of the signified. To put it differently, if an occurrence means that P in a natural sense (henceforth mean$_n$), then it is the case that P. This is called *the factivity principle* in Grice's terminology. A very commonly used example of the principle is the indicative relationship between 'red spots on the face' and 'having measles'. There is a lawful relation between red spots on the face and having measles. This lawful relationship is what constitutes the natural meaning of red spots. However, in some instances red spots might be caused by other factors. In such situations, according to Grice, the symptom of having red spots loses its natural meaning. In other words, the red spots on Tommy's face mean$_n$ that Tommy has measles only if Tommy really has measles (mean$_n$ = means in a natural sense). In natural meaning instances, the relationship between a sign and the entity that is signified is one of indication.

Non-natural meaning instances, as opposed to natural meaning ones, arise from non-natural signs, to wit, signs that are created via some sort of convention. For example, the commonly known hand gesture made by forming a V with two fingers means 'peace', and this meaning is acquired by convention. The conventional meaning of such a sign can change from one context to another. Besides these hand gestures, any natural language is also a result of conventions, according to Grice. The main characteristics of those non-natural meaning instances is that they can go wrong; when one sees the V sign formed by two fingers, the sign may refer to something other than peace[10], or when one hears the word 'dog' uttered, this does not necessarily mean that there is a dog in the vicinity. In other words, non-natural meaning instances do not obey the factivity principle. The relationship between a sign and the signified, in the case of non-natural meaning, is one of representation. As a result, as opposed to natural signs, non-natural signs have the capacity of misrepresenting.

Naturally, it is not enough just to propose the distinction between natural and non-natural meaning instances, one also needs to offer a set of criteria by which the naturalness and non-naturalness of a meaningful instance can be identified. For this purpose, Grice offers a twofold recognition test. For a sign X to mean r in a natural sense the following two criteria must be satisfied [Grice 1989, p.349]:

- It should be the case that 'X means that r' implies that r. (The factivity principle).
- It should be possible to rephrase the assertion of meaning in the form 'The fact that X means that r'. (The quotability principle).

If a meaningful instance does not satisfy either one or both of these criteria, then it is a case of non-natural meaning. The details of Grice's recognition test are not our main concern here. The only point is that such a test leads to a dualism of meaning . That is to say, a meaningful instance either falls under the category of natural meaning or it is a case

[10]In fact, in World War II the Allied soldiers used the V sign for victory. An ironic transformation happened during sixties when the beat movement used the sign for peace in demonstrations against the Vietnam War.

of non-natural meaning. This dualism of meaning[11] is what Dretske borrowed from Grice besides the original distinction itself.

The overarching goal of Dretske's project is to naturalize mental content. So, applying Grice's dualism of meaning, he starts with the notion of natural meaning for defining his informational content. Since natural meaning instances obey the factivity principle, he puts the strong constraint of implying truth on this notion. This is the fundamental motivation behind assigning unity to conditional probabilities (it is worth stating that Dretske himself did not mention this connection between assigning unity to conditional probabilities and the factivity of natural meaning in his book). Then Dretske tries to explain mental content, which obviously falls under the category of non-natural meaning since it does not obey the factivity principle, in terms of informational content without introducing any intermediary category (or overlapping category for that matter). This is the biggest handicap, i.e. not having any intermediary or overlapping category between natural and non-natural meaning, that Dretske creates for himself.[12]

The dualistic analysis of meaningful instances proves to be crucial for Dretske's project, so it is only natural to ask whether it is satisfactory. A close scrutiny shows that there are two odd consequences of this dualistic approach. The first one is that it provides only particular natural meaning instances, to wit, it lacks a type-level meaning for natural signs. Secondly, the distinction is not exhaustive. The dualistic natural versus non-natural distinction is not capable of capturing some essential meaningful instances. Let me delve into these two consequences one by one.

---

[11]Although Grice said that he did not want to offer a very sharp distinction, his expectation was that in most cases the instance would fall under either of these categories (an exclusive disjunction), and moreover the structure of the test itself implies a dualism of meaning. For a detailed discussion of this point, please see Denkel [1995].

[12]As discussed in Chapter 2, Dretske realized his mistake in 1986, and tried to fix it by introducing a third category, functional meaning. However, this did not solve his problems either, because his new suggestion still carried a particularistic approach to meaning. Instead of two mutually exclusive categories, he ended up with three mutually exclusive categories. The idea behind his 1981 and 1986 suggestions is essentially the same, and it misses the importance of a continuum among different meaning categories.

Following in Grice's footprints, Dretske endorses the factivity principle of natural meaning, and this leads him to a particularistic account of meaning. Natural meaning is accepted only in particular instances. Let's take the red spots and measles example. The particularistic account of meaning implies that there is no type-associated meaning of the symptom of having red spots on the face. We cannot talk about the meaning of having red spots on the face in a general sense, according to Dretske. Whenever we talk about this symptom and its relation to having measles we should be talking about particular instances in which the person really has measles. He clearly mentions this in his *Misrepresentation* article.

> In speaking of natural meaning I should always be understood as referring to particular events, states or conditions: this truck, those clouds, and that smoke. [Dretske 1994, p.159]

Now imagine that you are looking at two people, Tommy and Alice, and both have red spots on their faces, but only Tommy has measles. The following three statements about this situation are compatible in the dualism of meaning that Dretske endorses.

(1) The red spots on Alice's face do not mean$_n$ that she has measles.
(2) The red spots on Tommy's face mean$_n$ that he has measles.
(3) Although the red spots on Tommy's face mean measles, having red spots as such does not mean measles.

Since one cannot speak of natural meaning instances in a general sense, there is no way of mentioning the connection between having red spots on the face and having measles in Tommy's case unless one wants to claim that the connection between red spots and measles is a result of non-natural meaning, i.e. some sort of convention. Not only is this an odd result, but to accept all these three statements together, to say the least, is counter-intuitive.

The second unacceptable result of the dualism of meaning is that some important cases of meanings are left out because they neither qualify for the category of natural meaning nor for the non-natural meaning category. Denkel, in his *Reality & Meaning*, examines several examples which the Gricean dualistic distinction fails to capture. The following

three examples that I borrow from Denkel are sufficient to prove the point [Denkel 1995, pp.186-190].

- A)That she has these spots on her skin means that she has the measles
- B)The hair erection of this cat means that the animal is scared.
- C)The bark of that vervet monkey means that an eagle is approaching.

Intuitively, we are inclined to consider these three cases as cases of natural meaning. However, once Grice's twofold recognition test is applied to these cases, the situation becomes problematic since none of these meaningful statements satisfies the factivity principle. As mentioned above, in the Gricean framework and thereby in Dretske's understanding of natural meaning, it seems perfectly acceptable to say that 'These spots mean measles, but actually she has not got the measles'. In a similar manner, the following are also acceptable. 'The hair erection of this cat means that the animal is scared, but she actually is not scared since the erection is caused by the vet's chemical shot' and 'The bark of that vervet monkey means that an eagle is approaching, but there is no eagle in the vicinity since the monkey is barking for the purpose of play with her friends'.

Since the above examples fail to pass the recognition test for natural meaning, they will have to be categorized as non-natural meaning instances. This option, however, is not acceptable either since there is no conventional apparatus involved in any of them. The result is that Grice's dualistic approach is not capable of explaining these three meaningful instances. Moreover, the list of such examples could be increased by using examples from some other areas such as instinctive human communication. Hence, the dualistic approach is impoverished in terms of explaining meaningful instances.

These two unacceptable consequences are sufficient to reject the dualistic reading of the Gricean distinction, but they do not discredit the original distinction itself. The original distinction between natural and non-natural meaning instances points out an essential difference between two types of meaning. So, the original distinction must be preserved by rejecting the dualistic approach. A continuum between natural meaning instances and non-natural ones, which allows for the existence of intermediary and overlapping categories,

seems to be a natural solution for the problem in hand. Such a continuous reading will keep the valuable distinction without falling into the trap of dualistic two mutually exclusive poles.

It is clearly true that there are some natural signs which do their job, i.e. refer to their signified, without any human interference, and moreover they obey the factivity principle. It is also true that there are some communication systems where the relationship between a sign and its signified is determined completely by conventions. However, it is equally true that there are several instances where neither the former nor the latter category seems to be a perfect fit. Such instances seem to be more of a mix of these two categories. The bark of a vervet monkey is a good example for such a mix. The meaning of the bark is nowhere close to being a result of convention, but on the other hand it does not fully qualify for being an instance of natural meaning either. Hence, it seems very intuitive to claim that it is a mix of natural and non-natural meaning types. The difference between the bark of a vervet monkey and a natural meaning instance seems to be the weak causal link between the bark and the existence of an eagle in the vicinity compared to the causal link between a natural sign and its signification, for example the causal link between an oasis and the existence of water in the vicinity. On the other hand, the causal link between the vervet monkey's bark and its signification, i.e. the existence of an eagle in the vicinity, seems to be much stronger than the causal link between the V sign and the idea of peace. Thus, in terms of the strength of the causal link between a sign and its signification, natural meaning instances have the strongest ones whereas non-natural meaning instances have the weakest ones. This difference gives us a criterion by which one can identify the location of a meaningful instance on the continuous line between natural and non-natural meaning instances. This idea could also be expressed in terms of the strength of the lawful regularity between a sign and its signification instead of appealing to causality. The strongest lawful regularity happens in the case of fully natural meaning instances whereas the weakest ones happen between fully non-natural meaning instances. Figure 3 summarizes this continuous reading of the original Gricean distinction. As shown in the figure, there could be several different intermediary categories since the strength of lawful regularity between a sign and

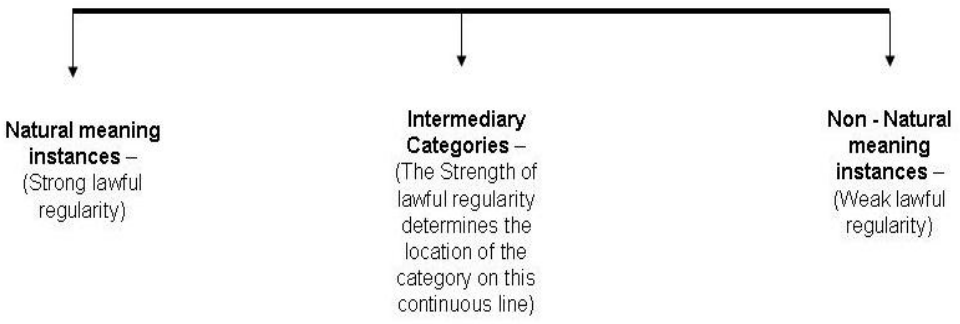

FIGURE 3. The Continuous Reading of the Gricean Distinction

its signification can vary on a wide range. Some examples of the intermediary categories are the following: functions based on natural mechanisms and animal communicative behavior. This continuous reading seems to be a better understanding of the Gricean distinction than the original dualistic reading which led Dretske to the trap of assigning unity to conditional probabilities. Within the continuous reading, the conditional probability of a signified given the sign varies depending on the location of the meaningful instance between two ends of the line of meaning. In the case of strict natural meaning instances, the conditional probability of the signified given a sign is one. One example is helpful here.

P('the presence of water in the vicinity' | 'an oasis') = 1

It is clear that Dretske has such instances in mind for the definition of informational content. The value of conditional probabilities starts decreasing as soon as one leaves the strict cases of natural meaning instances. One example from the literature on biological functions is the following:

P('the presence of a fly' | 'a frog's sticking out its tongue') < 1

It is less than one, because, as it is well known by now, frogs stick out their tongues even when they see fly-like black dots.

It seems that the Dretskean framework would have more resources to deal with problems like misrepresentation if a continuous reading was accepted. On the other hand, the

continuous reading does not lead to unacceptable consequences as in the case of the dualistic reading. Thus, I claim by following Denkel's footprints, Dretske's original theory must be revised by providing a continuous reading between natural and non-natural meaning. Dretske, in his original theory, tried to reduce mental content to natural meaning instances via the definition for the notion of informational content. Instead of treating mental content in that way, one needs to identify the location of mental content on the continuum line of meaning, and then its connection to natural meaning instances.

3. A True Probabilistic Definition

Dretske's three arguments for assigning unity to conditional probabilities have been analyzed in Section 1. As the analysis shows, none of the arguments was strong enough to imply Dretske's strict constraint. Hence, there is no argument against assigning values less than one to conditional probabilities in the definition of the informational content of a signal. Moreover, the previous section, Section 2, identified the fundamental motivation, the Gricean distinction, behind assigning unity to conditional probabilities in Dretske's framework. The Gricean distinction is a good starting point for naturalizing mental content. However, as it is defended in the previous section, the dualistic reading of the distinction is the only rationale behind Dretske's strict constraint. There is a better reading of the Gricean distinction: the continuum between natural and non-natural meaning categories. Not only is the idea of continuum consistent with assigning values less than one to conditional probabilities between signs and their significations, but also assigning values less than one is necessitated by the idea of continuum. That is to say, conditional probabilities range from one to smaller values depending on the lawful regularity (or statistical dependence) between the sign and its signification (or indication). This approach, the ordinal ranking approach, benefits from the Gricean distinction without paying the price of Dretske's strict constraint. In short, the previous sections serve as an existence proof for the ordinal ranking approach in defining the notion of informational content. Such an existence proof is not sufficient unless a formal definition of the ordinal ranking approach is provided. This section aims to achieve this goal.

The simple idea behind the ordinal ranking approach is the following. Take the given sign and the conditional probabilities of possible contents, and then pick the content which has the highest conditional probability. The idea is simple but to state it formally is an intellectually challenging task. As mentioned above, the main challenge is to find a tool by which the set of all possible contents could be identified. One notion that Lehrer uses in his 1971 article, briefly mentioned above, is helpful for this purpose.

Probabilistic frameworks lead to some paradoxical results; the lottery paradox is one realization of these. If any probabilistic threshold were enough for knowledge claims, then the probabilistic evidence for not winning a lottery would be a good candidate. Assuming that there is such a probabilistic threshold, a person who bought a lottery ticket 'knows' that he is not going to win. However, if this is true for one person, then it is true for everyone that has bought a ticket for the lottery. That is to say, everybody who has bought a ticket 'knows' the denial of the claim that s/he will win. This is paradoxical, because one person always wins (at least in an honest lottery). The paradoxical result arises because the set that contains the denial of the hypothesis 'I am going to win' for every ticket holder is an inconsistent set. Let me use a little bit of formalism.

h_1 = 'I will win' for person 1.

h_2 = 'I will win" for person 2.

h_3 = 'I will win' for person 3.

And so on.

Hence, the denial set is {not h_1, not h_2, not $h_3 \ldots$}. The inconsistency of this set, which is the basis of the lottery paradox, is a special inconsistency; it is minimal. It is inconsistent, but not any of its proper subsets. For example, the set that contains not h_1, not h_2 and not h_3 is consistent, and this is not true for the entire set. The inconsistency of the entire set is called **minimal inconsistency**.[13] Lehrer uses this notion in his probabilistic framework in order to avoid the lottery paradox and in order to prove that there is no arbitrary threshold for probabilistic knowledge. The other property of this set is to be able to list all

[13]In algebraic logic, many concepts come in pairs. Minimal inconsistency is the dual of maximal consistency.

possibilities indirectly. It does not list all h_is; rather it lists the denial of each h_i. Not plain inconsistency, but only minimal inconsistency guarantees the possible reach to all possible hypotheses, though indirectly via their negations. Here is a formal definition of minimal inconsistency, introducing the locution 'MIS' to refer to such a set.

> **MIS:** A set S is MIS if and only if S is inconsistent and every proper subset of S is consistent.

This notion is crucial for the definition of informational content that is being developed. However, it needs to be expanded by including the notion of background knowledge. Even in real life applications, the informational content of any signal is understood within the context of the background knowledge of the system that receives the signal. Here is the extended definition of minimal consistency with respect to some background knowledge - let b stand for that background knowledge.

> **MIS$_b$:** A set S is minimally inconsistent with respect to background knowledge b (MIS$_b$) if and only if the set consisting of b and the members of S is minimally inconsistent.

With this notion, all possible hypotheses, in other words all possible contents, are indirectly available to us. The indirectness means that only the denial of each possible content is included in the set. The idea of the ordinal ranking approach is to find out the one possibility that gives us the highest conditional probability. The indirectness of the notion of minimal inconsistency avoids just going after the highest number. Let me explain this with a toy example. Assume that there is an informational signal - r - that is received by an organism. And there are three possible alternatives as the background knowledge that can be attached as the content of this signal - say COW, DOG and CAT. The set that contains the denial of each of these is minimally inconsistent with respect to the background knowledge of the organism.

Toy = { not COW, not DOG, not CAT}

This is the only set that is minimally inconsistent. The set that contains the positive alternatives, {COW, DOG, CAT} is inconsistent but not minimally, because some of its

proper subsets are also inconsistent. Therefore, there is no principled way of choosing {COW, DOG, CAT} over {COW, DOG} since both of them are inconsistent. This is the main rationale for using minimal inconsistency. So, we know that we have to use the set that includes negative hypotheses (the set Toy in our example).

In the example, we want to be able to pick the most likely option among COW, DOG or CAT. From the set Toy, if we were to choose the option with the highest probability, then we would end up with the least likely option and that would be a mistake - in fact Lehrer in his original article makes that mistake.

Toy = { not COW, not DOG, not CAT}

Assume that the list of the conditional probabilities of the members of Toy given the signal, r, is the following.

P(not COW | r) = 0.6

P(not DOG | r) = 0.4

P(not CAT | r) = 0.2

The highest number in the list is 0.6, and this gives us the least likely option, because the conditional probability of COW given r, P(COW | r), is equal to (1 - P(not COW | r)), i.e. 0.4. When this simple calculation is applied to the other two numbers in the list, it will be seen that the most likely option is the one with the lowest number in the list, which is CAT in our example. Therefore, we have to choose the lowest number, or in other words we have to choose the conditional probability which is lower than all the others. Once again, to pick the highest number given a minimally inconsistent set would be an unfortunate mistake that Lehrer makes in his otherwise excellent article.

The notion of minimal inconsistency and our simple example give us enough ammunition for defining the notion of informational content.

Informational Content (IC): A signal r carries the information that h given the background knowledge b if and only if for any set which is minimally inconsistent given b (MIS_b) and includes $\sim h$ (where $\sim h$ is not h), then for every other member, k, of this set, Pr ($\sim h$ | r) is less than Pr (k | r).

In plain language, the formal definition states that a signal carries the information that is the most likely among all available alternatives to the system. The available alternatives are stored in the memory mechanism of the system. In a sense, the main job is done by the memory of the system (this point will be discussed in Chapter 6 in detail). This is a problematic claim since it is not difficult to think of an information receiving system (or even a perceptual system) that has no memory mechanism. Moreover, behavioral and neurobiological research on perception provides strong evidence for early perceptual mechanisms (which are information receiving mechanisms) that don't employ any memory mechanism. Given these, the above formal definition seems to have a flaw. However, this is not exactly correct. The above definition is a general one; it has different special realizations. One of them is where no memory related phenomenon is involved, and this realization is consistent with or rather is implied by the strict Xerox Principle. As mentioned in the second section, Dretske's strict Xerox Principle that implies unity for conditional probabilities is not completely wrong; rather it applies to a limited set where the informational flow forms a Markov chain. There is such a limited set in the case of perceptual mechanisms as well. The most basic and fundamental interaction between the external state of affairs and a perceptual system forms such a set. I will explain this a little bit more in the following paragraphs; for now let me formally define the special realization of the definition of informational content where no memory is involved, and is consistent with the strict Xerox principle. For this purpose, imagine that $P(\sim h|r)$ in the above definition is zero, which is theoretically possible. If this is the case, then the system does not need to compare it with any other alternatives, since all other conditional probabilities are greater than zero.[14] Since P(h | r) = 1 - Pr ($\sim h$ | r), Pr (h — r) is one. That is to say, h is implied by r. These facts give us the special realization of IC and let me introduce the locution 'ICs' for 'informational content synchronic' for this special realization. I call it synchronic since it is supposed to show the informational change at the time of interaction without referring to any memory mechanism, i.e. without any reference to past experiences.

[14]This is the very first of axiom of the probability calculus.

Informational Content Synchronic(ICs): A signal r carries the information that h at the synchronic level if and only if P($\sim h$ | r) is equal to zero (where $\sim h$ is not h).

In a sense, the interaction between the information receiving system (for example human perceptual system) and the information producing source (e.g. the external world) at the synchronic level forms a perfect communication.[15] The possibility of such a perfect communication was proven by Shannon in his original formulation of the mathematical theory of communication [Shannon & Weaver 1980, p.71], and his proof of this claim is known as the fundamental theorem of the mathematical theory of communication. Such a level where perfect communication occurs exists almost in any information receiving system. In the case of computers, this is the level where a particular electric current is transferred from one switch to another.

It should be noted that the synchronic level realization of the notion of informational content is essentially identical with Dretske's original definition. So, one needs to admit that Dretske's original definition is not wrong; rather it is limited to a more basic level. The main problem with the original formulation is that Dretske overgeneralizes it, and applies it to the levels where it does not apply. As a result, his original framework fell short of utilizing the benefits of the mathematical theory of communication and appreciating the importance of the memory mechanisms of an information receiving system. In 1981, Dretske was well aware of the importance of the memory mechanisms of a system, and he tried to incorporate that notion in his distinction between the learning period and the retrieval period of a system. He was right in realizing the importance of the memory mechanism; the only mistake was the way he implemented the notion. He implemented the notion with an unprincipled distinction, the distinction between learning and retrieval periods of an organism. After receiving immense criticism for his unprincipled distinction, he took his teleosemantical turn in 1986. With this turn, however, not only did he give up the unprincipled distinction but he also gave up the notion of memory in his framework. One

[15]This does not mean that there is no noise. Noise can theoretically be excluded.

cannot find a better place for using the idiom 'throwing the baby with the bath water'. In a sense, in my dissertation, I attempt to rescue the baby.

CHAPTER 5

An Alternative Concept: Mutual Information

I provided a probabilistic framework for informational content in terms of inverse conditional probabilities and minimally inconsistent sets in the previous chapter. However, in Chapter 3, we have also encountered a problem that the use of inverse conditional probabilities faces: the probability interpretation problem. In that chapter, I offered a solution for this problem as well. Given this solution and the benefits of using inverse conditional probabilities (taking the animal's perspective as discussed in Chapter 3), I insist on using inverse conditional probabilities in the definition. On the other hand, after Dretske's 1981 attempt at defining informational content in terms of conditional probabilities, several people claimed that the mathematical theory of communication provides us better notions that do not necessarily appeal to inverse conditional probabilities. The main notion that has been proposed is the notion of mutual information. Grandy [1987], Harms [1998], Usher [2001], Eliasmith [2000, 2005] are just some examples of such an attempt. Given the prevalence of such attempts, it is only natural to ask whether using the notion of mutual information could be better instead of insisting on the use of inverse conditional probabilities in defining informational content. In this section, I aim to answer that question.

The notion of mutual information between two variables is simply a similarity measure that Shannon's theory of communication uses for information flow [Shannon & Weaver 1980]. It has properties that are useful for defining informational content and mental content. However, the main problem with that notion is that it is symmetric. That is to say, the mutual information between r and s is identical to the mutual information between s and r. When this notion is used for defining mental content, it encounters the problems similar to the ones that afflict the resemblance theory of representation. To give a very simple example, my mental representation DOG represents a dog out there in the world, but the

dog out there does not represent my mental representation. In other words, the notions of mental content and mental representation are one way relations. Mutual information being a symmetric property fails to account for this property of mental representations. For this very reason, the resemblance theory is not a successful candidate for explaining away mental content and representation [Fodor 1992; Cummins 1991].

Despite the fact that mutual information fails to be a good candidate for defining informational and mental content, the search for a better notion within the repertoire of the mathematical theory of communication is not futile. There is another notion which has all the positive properties of mutual information without being symmetric. It is the Kullback - Leibler divergence measure [Cover 1991, p.18]. Thus it is possible to define informational content without facing the difficulties that come with the use of inverse conditional probabilities. Yet, I still insist on using conditional probabilities. The main reason for my insistence is that Kullback - Leibler divergence does not provide a clear way of distinguishing the synchronic level from the general levels, a distinction that I provided in the previous chapter. This distinction is essential for this dissertation. In the second section of this chapter, I provide a short analysis of Kullback - Leibler divergence measure and my reasons for not incorporating it into the definition of informational content.

1. Mutual Information

Since the first edition of Shannon & Weaver's seminal article, *The Mathematical Theory of Communication*, both philosophers and psychologists have had their eyes on the notions that Shannon develops for their own purposes. They realized the potential value of notions such as information, entropy and channels for solving philosophical and psychological problems. After all, the relation between the human mind and the external world is one of communication, and Shannon's formalism has proven to have a high explanatory power in the notions of communication channels and information transmission. After the initial reaction, philosophers realized that there are fundamental differences between Shannon's information and the notion that they need for their own purposes. Shannon's project was to formalize the best way of coding and encoding messages for communication purposes.

Given these engineering purposes, he had to work at an abstraction level where the content of a signal did not matter. After all, he needed a theory that could be applied to any content that might be communicated. A little bit of detail about the main problem that Shannon wanted to solve is useful for our purposes.

Shannon's main question was the following: given a set of possible states, what is the expected surprisal value of a particular state that belongs to the set of all possible states. More formally, What is the expected value of a random r_i where r_i is a member of $S = \{r_1, r_2, r_n\}$? He started out with three basic intuitions:

- The expected value should depend only on the probability of r_i, not on the content of r_i
- Expected surprise should be a kind of expected value
- The expected surprisal value of an r_i should increase when r_is become more equiprobable.

The last intuition is similar to the case of a fair and unfair coin. The result of a toss of a fair coin is more surprising than that of an unfair coin. Surprisingly enough, the only set of functions that satisfy these three intuitions is the set of entropy functions of thermodynamics[1]. The very first of these three basic intuitions that Shannon had implies that his theory is not about the content of a signal, rather it is about the amount of information that a signal or a probability distribution for a set of states has. This is the point where Shannon's and philosophers' interests diverge. Philosophers are interested in identifying the content of a signal where the signal may be a linguistic or mental entity. This divergence led philosophers to search for a more suitable notion of information [Bar-Hillel 1955; Hintikka 1970]. Dretske's 1981 attempt to use the tools of Shannon's theory also falls under this category.

Dretske, in his 1981 book, clearly states that Shannon's theory is not very useful for epistemology or philosophy of mind, because it is about the average information that a set

[1]Several people claimed that this connection between the entropy of thermodynamics and the measure for expected surprisal value (information) points out some deep metaphysical connections [Wiener 1961; Wheeler 1994; Chalmers 1996; Brooks & Wiley 1988].

of messages has whereas epistemology and philosophy of mind are concerned about whether a person knows (or acquires) a particular fact on the basis of a particular signal. In other words, philosophical issues require the notion of information as having a specific content, not just the amount of information that the signal carries. Despite these diverging interests, Dretske believes that some notions of Shannon's theory could be a starting point for solving philosophical problems. He borrows the notion of entropy of a signal from Shannon and develops his own theory of information. So far, we have seen both positive and negative sides of Dretske's theory. After Dretske's attempt, however, several people have claimed that Shannon's theory could be more useful for philosophical purposes than Dretske claims. Grandy [1987] provides an information theoretic approach based on Shannon's mutual information, and he claims that a proper use of mutual information could serve as a basis for an ecological and naturalized epistemology. Harms [1998] claims that, on the other hand, mutual information provides an appropriate measure of tracking efficiency for the naturalistic epistemologist, and this measure of epistemic success is independent of semantic maps and payoff structures. Usher [2001] proposes a naturalistic schema of primitive conceptual representations using the statistical measure of mutual information. Eliasmith [2000] in his unpublished dissertation uses the same notion for defining both mental and neural content, and constructs a neurocomputational theory of referential content. In order to see how the notion of mutual information plays out in regards to philosophical problems, it is useful to discuss these attempts one by one in more detail.

1.1. Grandy's Ecological and Naturalized Epistemology. Grandy's ultimate aim is to flesh out Quine's notion of 'epistemology naturalized'. He agrees with Quine on seeing epistemology (and philosophy of mind to an extent) as a part of psychology, but he wants to go further and identify the branch of psychology that epistemology and philosophy of mind should belong to. Let me state Grandy's conclusion in his own words at the outset.

> My own view, just a little less sketchy than Quine's, is that epistemology should be seen as a part of ecological psychology as developed by Gibson,

> but that the concept of information involved should be identified with that of Shannon and Weaver (p.199)

The central concept that Grandy borrows from Shannon's theory is *average mutual information*. This notion is defined between two sets of alternatives (more formally between two random variables). Grandy's explanation of the notion is as follows. Let B_i and C_j be two sets of alternatives (where $0 < i < n$ and $0 < j < m$ for finite n and m), and these sets are disjoint and exhaustive. Furthermore, assume that there is a well-defined probability distribution for the joint probabilities of B_is and C_js. For any B_i and C_j, if they are correlated, i.e. $P(B_i|C_j) > P(B_i)$, then C_j conveys some information about Bi and vice versa. The average amount of information that these possibilities convey about each other is the mutual information between the sets of B and C. This amount is calculated with the following formula.

$$\sum_i \sum_j P(B_i \& C_j) \times \log \frac{P(B_i \& C_j)}{P(B_i) \times P(C_j)}$$

Mutual information, Grandy claims, is consistent with the Gibsonian idea of ecological perception. In Gibson's ecological psychology, senses are the systems that acquire information about the environment [Gibson 1966]. He is also clear on emphasizing that his notion of information is different than Shannon's notion of information. However, Gibson does not specify any specific reason for the difference. Grandy identifies potential reasons for rejecting Shannon's notion in Gibson's writings. Moreover, he shows that the notion of mutual information has the potential of satisfying these concerns. Grandy identifies four possible concerns. In Gibsonian psychology, the senses are actively involved in constructing perception. In other words, the perceiving agent is not just a passive receiver; rather the agent is actively involved in picking up information about the environment. On the other hand, the common way of presenting the information theory is in terms of a source sending a signal to a passive receiver. This is a probable reason why Gibson's notion of information is claimed to be different than Shannon's notion. However, adds Grandy, the notion of mutual information is perfectly symmetric. In the formula stated above, it does not matter whether B or C is sending the signals; the mutual information between these two sets of

alternatives is the same in either direction. Hence, the concern about passivity vanishes once mutual information takes its place on the stage.

The second possible concern that Grandy identifies is about channels. Shannon's theory, as a result of being intrinsically an engineering approach for communication, focuses on channel conditions between the source and the receiver. Again, Gibsonian psychology's emphasis is on the active perceiving organism not on the channel conditions between the stimulus and the organism. This might be another reason for Gibson to differentiate his notion of information from Shannon's notion. However, as Grandy rightly points out, in the formal definition of mutual information there is no reference to channel conditions; it only requires identifying relevant conditional probabilities. Hence, this second possible concern does not have merit, either.

The third possible concern is about digitalization. The commonly known version of Shannon's theory deals with discrete signals. Gibson is interested in continuous quantities of the environment such as distance and velocity. If discreteness were to be the only version of Shannon's theory, then Gibson would have had a very good reason for using a different notion of information. However, Shannon's theory could easily be generalized to continuous variables and channels as well. In fact, the second half of the seminal article is devoted to generalizing informational properties of discrete variables and channels to continuous ones. The formulas and theorems of the discrete version are true of the continuous version as well with only one basic difference: the sigma (summation over discrete states) is replaced by a stretched S (summation over continuous states: the integral sum).

Specification is the fourth possible concern that Grandy discusses. One of the essential features of Gibsonian psychology is that the proximal stimulus, i.e. the incident optic array on the retina, carries enough information to specify the distal stimulus (the stimulus in the environment). In other words, specification of the distal stimulus is perfect in Gibsonian psychology. However, Shannon's theory allows information to be nonzero even where less than perfect situations obtain. This might be another potential concern for Gibson. Grandy offers a solution for this possible divergence as well. He says that it is consistent with Shannon's theory to have an extra assumption that the mutual information

between distal and proximal stimulus always equals the entropy of the distal stimulus. This would guarantee that the proximal stimulus always perfectly specifies the distal stimulus. After analyzing these four possible concerns, Grandy concludes that Shannon's notion of information is perfectly consistent with the ecological theory of perception, and hence both Gibsonian ecological psychology and Shannon's theory of communication could benefit from each other.

As mentioned above, Dretske claims to have a notion of information different from Shannon's, because Shannon's notion of information is not suitable for epistemological purposes. Grandy argues that Dretske's claim is true for a naturalized version of traditional epistemology, but not for an ecologically motivated epistemology. To justify his claim, Grandy uses frogs' ability to catch bugs when they are present as an example which is very prevalent in the literature. He considers several different scenarios where the correlation between a bug's presence in the environment and the catching behavior of a frog varies from perfect to less than perfect cases. The details of these scenarios will take us far from our subject. However, Grandy's conclusion is important for us. To put it briefly, his conclusion is that mutual information together with the commonly accepted notion of expected utility perfectly matches our intuitions about which one of those scenarios is preferable ecologically. Moreover, he shows that Dretske's theory does not match our ecological intuitions where less than perfect situations occur. Thus, he concludes, Dretske's theory is not an ecologically motivated information theory, but it is possible to offer an ecologically suitable form of information theory which is loyal to Shannon's notion of information.

1.2. Harms's Tracking Efficiency Measure and Payoffs. Harms [1998], in his *The Use of Information Theory in Epistemology*, undertakes an ambitious task which has two main components.

- To identify the relevant measure of information for tracking efficiency of organisms
- To flesh out the relationship between the information measure and payoff structures.

For the first part of his task, he offers mutual information as the right tracking efficiency measure. For the second part, he shows that mutual information is independent of payoff structures.

Harms's claim about mutual information is very close, if not identical, to Grandy's position in regards to the value of Shannon's original theory for philosophical problems. Both think that the gap between Shannon's information and philosophically relevant information is not as big as some others, like Dretske, think. Here is what Harms says about the nature of this gap.

> It is one thing to calculate the accuracy of sending and receiving signals, it is another thing entirely to say what those messages are about, or what it *means to understand* them. Consequently one might think that since the notion is not semantic, it must be syntactic or structural. The dichotomy is false, however. What communication theory offers is a concept of information founded on a probabilistic measure of uncertainty. However, even respecting that information theory does not presume to quantify or explain meaning, there remains the possibility that the information theoretic notion of information can be applied to semantic problems. [Harms 1998, p.481]

The other task that Harms tries to accomplish is to flesh out the connection between Shannon's notion of information and payoff structures. This is a very important task, because since the mid 80s the dominating trend in the literature is to appeal to functions for explaining mental content. Millikan's criticisms [Millikan 1989] of Dretske's information theoretic approach and correlation semantics paved the way for moving away from the use of information theory towards an evolution-inspired explanation based on biological functions. Millikan's main worry is that information theoretic approaches put their emphasis on representation production. However, says Millikan, the crucial feature of an organism is representation consumption. This is what information theoretic approaches miss in explaining away the properties of mental representation and content. In fact, as a result of

Millikan's worries Dretske himself gave up on his information theoretic approach in 1986 and has been defending a teleosemantical approach similar to Millikan's (of course with some differences). Given the prevalence of teleosemantical approaches based on notions like utility, payoffs and functions, Harms's effort for fleshing out the relationship between mutual information and payoff structures is an invaluable one.

Harms provides two different analyses for the relationship between payoffs and information. The first one is based on randomly chosen probability distributions and the second one is based on the process of optimizing response mechanisms. The details of both analyses can be found in his article. For our current purposes, suffice it to state his conclusions without giving further details. The conclusion for the first type of analysis is that high information does not always guarantee high payoffs, but 'information still places upper and lower bounds on payoffs.'[Harms 1998, p.489] This type of analysis is based on randomly chosen probability distributions, and it may not be very appropriate for teleosemantical concerns since teleosemantical approaches uses evolution as their paradigmatic framework. Evolution is an optimizing process. Hence, using an optimization for identifying the relationship between payoff structures and information will be more in line with evolution based explanations. Given such concerns, Harms borrows a notion of optimization from microeconomics: Pareto optimization. Pareto optimization is a process by which the utility for a part of a system is increased while the utility for other parts are not decreased. In other words, there is some gain for some without having a loss for others is the main idea behind Pareto optimization. Harms's conclusion from applying Pareto optimization to the relationship between information and payoff structures is the following.

> What these results indicate is that in systems where responses to environmental states are adaptive, information will tend to be systematically, if not monotonically, increased as payoffs due to differentiating responses increase (p.497)

I find this result both interesting and groundbreaking, because if Harms is right, then appealing to teleosemantics does not gain us anything more than information theoretic approaches are capable of providing.

1.3. Usher-Eliasmith Line. Usher [2001] and Eliasmith [2000, 2005] both use the notion of mutual information for getting away from the stringent constraints of a causal theory of mental content. As we have seen in Chapter 1, causal theories cannot make room for misrepresentation cases because causal theories conflate the truth conditions of a mental entity with the conditions that determine the content of the entity. Usher claims that the statistical properties ingrained in Shannon's theory of information makes it possible to distinguish these two. He develops a statistical theory of referent on the basis of mutual information. Eliasmith, in a very similar line, defines mental content in terms of statistical dependencies, and uses mutual information as a measure of such statistical dependencies. Moreover, Eliasmith claims that such a measure solves the problem of neurosemantics, i.e. how a population of neurons acquires its content when activated. Each of these works deserves special attention and a detailed discussion. However, the scope limitations of this dissertation force us just to mention these ideas. The only purpose here is to use especially Eliasmith's work as an existence proof for the possibility of using mutual information for successfully describing the content problem in cognitive neuroscience.

In conclusion, Grandy's work shows us that an information theoretic approach loyal to Shannon's original theory has the potential of providing an ecological theory of knowledge and perception. Harms's work implies that the advantages of incorporating utility and pay-off structures into our theories of mental content are already ingrained in Shannon's notion of mutual information. The Usher and Eliasmith line shows the applicability of mutual information not only to mental content issues but also to neurosemantical ones. Thus, it is natural to wonder why we should not use mutual information in defining informational and mental content.

2. Mutual Information vs. Kullback-Leibler Divergence

Despite all its advantages, mutual information is not a good candidate for defining informational and mental content. As mentioned above, it is a symmetrical notion. That is to say, the mutual information between A and B are the same regardless of the direction of the relation between A and B. This turns out to be a problematic property for mental representations. The problem here is identical to one of the problems that the resemblance theory of representations faces. In Chapter 1, we have seen that the resemblance relation is a symmetrical one, but the representation relation is not. Thus, we concluded, together with philosophers like Fodor [1992] and Cummins [1991, 1996], that resemblance cannot explain representation. The same idea applies to the notion of mutual information as well. My mental entity DOG represents the existence of a dog out there in the world, but the dog does not represent my mental entity DOG. However, the mutual information between DOG and the dog are exactly the same as the one between the dog and DOG. That is to say, the notion of mutual information does not give us any tool for distinguishing the directionality of representation relation. I take this to be a good enough reason for rejecting mutual information as a plausible candidate for defining informational and mental content. However, we have seen several positive features attached to this very notion. These positive features stem from the statistical and mathematical properties of Shannon's theory of communication, but these positive features are not peculiar to mutual information. Shannon's theory has another concept which carries all the hallmarks of mutual information without being symmetrical. This notion is Kullback - Leibler divergence measure. In common presentations of information theory, it is called Kullback - Leibler distance measure. I think that this is a case of misnomer since the notion of distance is inherently symmetric but Kullback - Leibler measure is not. Thus, Kullback - Leibler divergence is a better name for the notion. The formula for calculating the divergence between two sets of alternatives (random variables more accurately) is the following.

Kullback - Leibler Divergence: $D(p \parallel q) = \sum_x p(x) \times \log \frac{p(x)}{q(x)}$

It is easy to show that the formula is not symmetric. It measures how the second variable diverges from the first one. The amount of divergence between the first and the second could be different from the one between the second and the first. Because of this property it is a better candidate for the representation relation. Moreover, it takes simple algebraic manipulations to show that when the Kullback - Leibler divergence between two variables decreases, the mutual information between them increases. Hence, it has all the mathematical properties associated with mutual information. It is possible to run Grandy's, Harms's, Usher's and Eliasmith's arguments by replacing the notion of mutual information in their definition with Kullback-Leibler distance. There will be no significant loss.

In short, if one wants to use a Shannon type notion in the definition of informational and mental content instead of conditional probabilities, Kullback - Leibler divergence is the one to use, not mutual information. Despite this fact, though, in the probabilistic account that I am developing in this dissertation I insist on using conditional probabilities instead of Kullback - Leibler divergence measure. As I mentioned above, the reason for that is the distinction between the synchronic and the general level that I briefly discussed in the previous section. Kullback - Leibler divergence measure does not have an upper limit and its lower limit is zero. The lower limit is achieved only when two variables involved are one and the same. Thus, it does not provide any property by which causal interactions can be separated from informational interactions. The synchronic level that I propose is supposed to capture causal interactions between external state of affairs and mental representations whereas the general level accounts for informational relations. The upper bound of conditional probabilities, i.e. the conditional probability of one, is suitable for the causal level with some idealizations. There is no such upper limit of Kullback - Leibler divergence. Thus, I use conditional probabilities in my definition of informational content for the purpose of capturing the distinction between the synchronic (causal) and the general (informational) levels. The following chapter clarifies this distinction in detail.

CHAPTER 6

Perception as Unconscious Inference

Simple truth miscall'd simplicity.

Sonnet 66, William Shakespeare

1. Two Levels of Informational Content

In this chapter, I offer a two level analysis of informational content of mental entities and clarify the relationship between these two different levels. The first level is where all the stimulus based information is acquired and passed onto the second level where the information is processed through the memory systems of the organism. The first level, because it is acting in the moment, is dubbed the synchronic level. The result of this process is an increase in the amount of information by utilizing past experiences of the organism. This is the second level and since it is defined through the memory systems of the organism, it is dubbed the diachronic level. I formalize the notion of informational content (and thereby mental content) for each of these levels. The combination of these two levels is what constitutes perception (and thereby mental representation). Given this framework, the process that is constituted by these two levels is a crucial point of the theory that I offer. In order to characterize this interaction, I utilize an idea that dates back to 11^{th} century: perception as unconscious inference. In a nutshell, any instance of perception is an *inference* based on both the stimulus based information and the memory based knowledge. Despite the fact that the inference I defend is not syllogistic or deductive, for clarificatory purposes using the syllogistic terminology is useful here. The synchronic level, in a sense, provides the minor premise whereas the diachronic level provides the major premise of the perceptual inference. After formalizing the two levels of perceptual informational content, I introduce the idea of perception as unconscious inference with a brief historical survey. Following

that, I offer my solution to the problem of misrepresentation and discuss some possible objections to my solution. In short, this chapter is organized as follows: i) formalization of two different levels of informational content; ii) perception as unconscious inference; iii) the solution and three possible objections.

1.1. Formalization. In order to motivate the idea of two different levels of informational content, I would like to use a simple thought experiment. Let's imagine that we have the technological means of destroying all memory mechanisms of a person while keeping perceptual mechanisms intact. Also suppose that we have the ability to restore memory mechanisms in a gradual manner whenever we want to, and the process of restoring is so successful that we end up with exactly the original person. Whether or not we really have such technological means is not crucial here. In other words, such a situation may not be technologically possible, but it is for sure empirically possible[1]. Now, let's assume that I volunteer for testing these technological means. Scientists in a secret lab erase all of my memories, destroy my memory mechanisms, but they are nice enough to keep my sensory mechanisms in place. Let's call this new state of my existence *Hilmi the deformed* (HilmiD hereinafter).

HilmiD has normal perceptual mechanisms in terms of all sensation modalities: visual, auditory, tactile, gustatory and olfactory. He is able to process information from any stimulus, but lacks the ability to process this information further than his detection mechanisms allow. It seems to me that there is a striking similarity between HilmiD and the marine bacterium that we saw in Chapter 2. Our marine bacterium, which was lured to a tragic end, had only one sensory detection mechanism: magneto taxis. HilmiD has several sensory detection mechanisms, at least one for each sense modality. However, HilmiD, like the marine bacterium, does not have any further means of processing the information that is acquired through these sensory detection mechanisms. Dretske in his analysis concluded that the marine bacterium is not capable of misrepresentation. Things may go wrong for

[1]Here I refer to the distinction between logical possibility, empirical possibility and technological possibility. Such a thing may not be technologically possible, but it is empirically possible because there is no known natural law that contradicts with such a thought experiment

the bacterium, one example being our evil deception with a magnet placed in the direction opposite to the earth's magnetic field. Another one is some malfunction in its magneto taxis system or some 'natural' noise in the earth's magnetic field. In neither of these cases does the bacterium instantiate a case of misrepresentation. Dretske's reason for this conclusion is that the bacterium does not have the capacity of associative learning. It is not necessary to confine ourselves to the notion of learning. It would be safer to use a more general concept in order to explain away the marine bacterium's situation. The bacterium does not have any further means for processing the information that it acquires through its detection mechanism. This is why it does not have the capacity of misrepresentation despite the fact that things can easily go wrong in its natural habitat. In order to avoid misunderstanding, I need to add a disclaimer here. Since Dretske brought the marine bacterium example into the literature, our understanding of the marine bacterium magneto taxis system has improved significantly. The magneto taxis mechanism is not as simple as presented in Dretske's 1986 article. Our current state of knowledge may imply that the marine bacterium has a primitive form of representation and misrepresentation. In the theory that I develop here the ability of misrepresentation is a result of the amount of inference that goes into detection of geomagnetic field. If our scientific understanding of the marine bacterium shows us that there is an inferential reference to the previous states of the marine bacterium in its magneto taxis mechanism then it will have misrepresentation capacity. This is a purely empirical question. The theory that I offer here provides a general criterion for misrepresentation cases. Moreover, as we will see in Section 3 of this chapter, I defend a continuum between the synchronic level (indication) and the diachronic level (representation). Therefore, the marine bacterium may end up having some primitive form of misrepresentation which is very close to the indication level. For now though, in order to be consistent with Dretske's original presentation, I assume that the marine bacterium does not have any capacity in my thought experiment.

It is illuminating to compare HilmiD's situation with the marine bacterium. It is sure that HilmiD's sensory system is much more complicated than the marine bacterium's. HilmiD, as mentioned above, has at least five different sensory detection mechanisms, if

not more. However, it seems to me that the complexity of HilmiD's sensory system is nothing more than a quantitative difference which does not amount to any qualitative difference. In other words, compared to the bacterium's sensory system, HilmiD's case is nothing but more of the same. Theoretically, one could construct a system as complex as HilmiD's system by connecting millions of the bacterium's magneto taxis systems in a connectionist architecture [McLeod *et al.* 1998]. For doing that one does not need anything more than a first-generation connectionist architecture. The main characteristics of first-generation connectionist architectures is that the network works only on the basis of a given input at a time without making reference to any previous activation values [Clark 2000]. By applying such an architecture to millions of magneto taxis mechanisms, in principle, one could have a connectionist network that is capable of detecting several different stimuli or several different aspects of one stimulus. It is generally claimed that first-generation connectionist architectures have no capacity of misrepresentation[2]. HilmiD's case is similar to such a connectionist network built out of millions of marine bacteria. If this claim is right, then HilmiD, like the marine bacterium, does not have the capacity of misrepresentation. Obviously, things can go wrong for HilmiD, but such instances would not amount to being a genuine case of misrepresentation (or representation for that matter).

It is useful to analyze the types of things that can go wrong for HilmiD. First of all, HilmiD's sensory mechanisms can malfunction. For example, the visual detection mechanism might produce something that it is not supposed to produce. Secondly, there could be some noise in the stimulus that is presented to HilmiD's sensory mechanisms. The marine bacterium's magnetotaxis mechanism can as easily malfunction as HilmiD's sensory system. When it does, though, we would not count that as a case of misrepresentation. Since

[2]I need to add a disclaimer similar to the one stated for the marine bacterium example. Depending on our analysis of the first-generation connectionist architecture, they could end up having some form of representation and misrepresentation which is closer to the indication level. The level of misrepresentation, however, in the first-generation connectionist architectures should be less than the marine bacterium's misrepresentation level.

HilmiD's case is not qualitatively different than the marine bacterium's case, there is no good reason for counting malfunction cases as genuine misrepresentation cases either.

For the case of noisy situations, there are two possibilities: the noisy situation is either a product of nature or it is artificially created as in the case of the planned deception of the marine bacterium. In the mathematical theory of communication, the best way of avoiding miscommunication due to noisy or corrupted signals is redundancy. Instead of sending one message, your best bet is to send many more of the same signal. If one of the messages gets corrupted due to some external factor, then it would be easier to detect that corruption by double-checking with the other copies of the sent message. The receiver of the message is less affected by possible noise as the number of the copies of the message increases. In fact, this solution is not just the solution of the mathematical theory of communication. It is nature's solution as well. Olshausen & Field [2005] in their analysis of images found out that there is an immense amount of redundancy in any given natural image. So, Nature provides a solution for a problem that it itself causes. However, it is still possible that Nature's solution may not work in some instances. Such cases would be identical to artificially created noise situations. The magnet that we placed in the marine bacterium's habitat in order to lure it to tragedy constitutes noise with no solution provided. In such cases, the organism will get things wrong, but as discussed in the case of the bacterium, such instances do not count as genuine cases of misrepresentation. In short, things can go wrong for HilmiD, but he does not have the capacity of misrepresentation.

HilmiD's sensory system tells us something about normal human beings' sensory systems. There is a level in our perceptual organization where the stimulus based information is acquired and processed without making any reference to our memory systems. One could call this an early stage of perception. At this stage, our perceptual mechanisms act like HilmiD's sensory system. There could be malfunctions and/or noise, but there is no capacity of misrepresentation. In other words, this level is entirely *indicative*. The information that is acquired at this level indicates the stimulus which is the source of acquired information. Despite the fact that things can go wrong for the system at this level because of noise, it is

legitimate to exclude noise from the picture for idealization purposes[3]. Excluding noise in that way is common practice in the mathematical theory of communication as well. It is not that we deny the existence of noise; rather we accept its existence and exclude it for idealization purposes. Once this is done, then the idea of indication becomes much more apparent. The signal that is acquired from the stimulus carries all the information that is required to identify the source. I think that Dretske's original formulation of informational content, where he assigns unity to conditional probabilities, corresponds to this level. This level, again using HilmiD's case as a motivating idea, operates at the moment with no reference to past. At this synchronic level, as a result of excluding noise, the conditional probability of the ideal (true) content given the signal is one. The ideal content is the source that causes the signal itself. In the previous chapter, because of the notion of minimal consistency, we used negation of contents in our definition of informational content. Since the probability of the content that we want is one, the probability of the negation of that content is zero. These remarks take us to the formal definition of informational content at the synchronic level.

> **Informational Content Synchronic(ICs):** A signal r carries the information that h at the synchronic level if and only if Pr ($\sim h \mid$ r) is equal to zero (where $\sim h$ is not h).

As mentioned above, this definition is almost identical to Dretske's original definition. This is the level that Dretske has in mind in his definition of informational content. However, since he stops at this level without introducing another level, as I will do in the following

[3]Prof. Colin Allen raised a substantial objection against the idea of excluding noise. The objection is that to exclude noise requires knowing the function of the sensory mechanism in question. Therefore, the idea of excluding noise is circular. This is a powerful objection. I have two responses to this objection. The first draws from Olshausen & Field [2005]. Nature, wherever the noise exists, provides enough redundancy. Therefore, redundancy in the signal could be used in order to identify the noise. Knowing the function of the mechanism is not required. The second response comes from the idea of minimal inconsistency. When noise is added to the set of relevant alternatives, which needs to be a minimally inconsistent set, then the minimal inconsistency constraint would be violated. Thus, it is possible to identify and exclude noise without making any reference to the function of the sensory mechanism in question.

paragraphs, his theory ends up not being able to accommodate the possibility of misrepresentation. This does not imply that his original motivation is wrong. On the contrary, his original motivation about invoking the indication relation as the basis of informational content is accurate. Therefore, the synchronic level satisfies Dretske's original motivation. Another way of seeing this fact is through the Gricean notion of natural meaning. As discussed in Chapter 4, Dretske's hidden motivation for assigning unity to the conditional probability in the definition of informational content is to naturalize informational content through Grice's natural meaning instances. These instances are purely indicative and obey the factivity principle (i.e. if A indicates B, then B is the case).

In conclusion, our perceptual system starts out with the synchronic information (or borrowing Irvin Rock's terminology stimulus-based information) acquired from an external state of affairs. At this level, the relation between external states of affairs and the organism's mental states is one of indication, and hence purely naturalistic. However, this is not the whole story of what is going on in our perceptual system. The organism processes the information acquired at the synchronic level by further means of its memory systems. In order to clarify this further processing, let me use HilmiD's case by gradually transforming him to original Hilmi.

Now, imagine that we take HilmiD and by using our memory technology we gradually restore his original memories and memory capabilities. He would slowly become identical to the original Hilmi who has the capacity of misrepresentation. At which point of this gradual transformation HilmiD gains the capacity of misrepresentation is an important question, but not very crucial for our current discussion. For now, suffice it to say that the transformed Hilmi would have different capabilities than HilmiD. Original Hilmi has the ability to misrepresent things in the world. So, by the assumptions of the thought experiment, after restoring his memory systems, Hilmi would have the capacity of misrepresentation. The question is that what gives this capacity to Hilmi. The answer to this question is pretty obvious because of the way I constructed the thought experiment. It is the memory capabilities of original Hilmi. The stimulus-based information acquired from the synchronic level is assigned a specific content through descriptions, concepts and rules stored in the

memory system. This is an inferential process that happens unconsciously - the idea of perception as unconscious inference is discussed in detail after formalizing the second level informational content. The content that is assigned to the stimulus-based information is picked out of the relevant alternatives. The idea is simply to pick up the one which has the highest probability among the relevant alternatives. The relevant alternative that has the highest probability has the highest amount of mutual information as well (or the least amount of Kullbeck - Leibler divergence). A simple example is useful here. Suppose that there is an organism that has only two categories stored in its memory system: CAT and DOG. The organism is presented with a scene where there is a cat. The causal connection between the cat and the organism provides information about three properties of the cat: furry, four legged and has tail. This could be thought of as the synchronic level where the relation between the organism and the external world is one of indication. At the synchronic level, the organism has three mental entities that indicate FURRY, FOUR LEGGED, HAS TAIL respectively. It should be noted that I use semantically transparent properties for ease of presentation. It is very likely that the properties that function at the synchronic level are not semantically transparent. That is to say, those properties are more fundamental than any property that we can linguistically identify. They are even more basic than properties like edge, orientation, momentum. For the distinction between semantically transparent and non-transparent properties please see Clark [2000]. At the next level, the organism tries to infer the proper content. It has two possibilities, CAT and DOG. Among these two, the organism picks the one that has the highest probability given the information of FURRY, FOUR LEGGED and HAS TAIL. The process described here is the process of the ordinal ranking approach that is discussed in Chapter 4. Since the second level is about the categories, descriptions that are stored in the memory system of the organism, it is natural to call it the diachronic level. The main challenge at this level is to be able to list all the relevant alternatives. In our toy example, the organism has only two categories and both of them are relevant to the scenery that is presented. So, listing relevant alternatives does not present itself as a challenge in the toy example. However, the list of relevant alternatives for complex organisms like human beings is rather challenging. Chapter 4 provided a formal

tool for overcoming this challenge: minimal inconsistency. The relevant alternatives are the members of the minimally inconsistent set available to the organism. The notion of minimal inconsistency works with the negations of relevant alternatives. Thus, the definition of informational content for the diachronic level needs to pick the one which has the lowest probability among the negations of relevant alternatives (please see Chapter 4 for details).

> **Informational Content Diachronic ICd:** A signal r carries the information that h given the background knowledge b if and only if for any set which is minimally inconsistent given b (MIS_b) and includes $\sim h$ (where $\sim h$ is not h), then for every other member k of this set, Pr ($\sim h$ | r) is less than Pr (k | r).

The process of selecting the one with the highest probability (the lowest probability for its negation) is an unconscious inference and because of its probabilistic character it is an inductive inference. As a result of this, things can go wrong in a substantial manner which gives us the possibility of genuine misrepresentation. One of the preliminaries stated in Chapter 1 is the source independent character of mental representation. Such an inductive inference provides a basis for satisfying this requirement. This is the gist of the solution that I offer for solving the problem of misrepresentation. The unconscious inference that happens at the diachronic level has two premises; one from the synchronic level which is the stimulus based information and the second is from the diachronic level which comes from the stored descriptions, categories and concepts. This inferential scheme is similar to the one offered by Rock [1975, 1983]. The characteristics of such an unconscious inference need to be specified for clarifying the relation between the synchronic and the diachronic levels. A brief historical survey of the idea of perception as unconscious inference is helpful for achieving the goal of clarification.

2. Perception as Unconscious Inference

The modern formulation of the idea that perception is mediated by unconscious inferences is due to the 19^{th} century scientist and philosopher Hermann Helmholtz. Helmholtz's overall aim, as detailed in Chapter 3, was to show that psychological states could be studied empirically. As a part of his overall project, he tried to explain the underlying perceptual

mechanisms of visual illusions [Helmholtz 1971, 1995]. The main conceptual tool that he used for this purpose was the idea of perception as unconscious inference. Helmholtz's conceptual tool has been a prevalent idea in the history of theories of visual perception. The idea that unnoticed, unconscious inferences underlie perception has been around since the time of Ptolemy. In the past millennium, this idea has been used by theoreticians like Alhazen, Descartes and Helmholtz. In the current literature, the most significant examples are Irvin Rock and, to an extent, Jerry Fodor. Hatfield, in his *Perception as Unconscious Inference* [Hatfield 2002], presents a very precise and informative historical survey of the ideas of these theoreticians. I shall follow his footprints and present a brief survey of those ideas.

2.1. Alhazen. Ptolemy [1996, ca.160] used the idea of unnoticed judgments for explaining perception about 11 centuries before Alhazen, but it was Alhazen who analyzed the underlying notion of judgment for the first time in detail [Alhazen 1989, ca.1030]. For Alhazen, any instance of perception beyond the passive apprehension of light and color necessitates some sort of judgmental/inferential process. Some examples of perception based on judgment/inference are recognition of color categories, perception of similarity and dissimilarity and distance perception. In such cases, judgment/inference works by comparing the passively apprehended stimulus with previous instances, and hence Alhazen emphasized the importance of previous learning in perceptual unconscious judgements.

Since Alhazen defined senses as passive recipients of stimuli signals, the task of judgment could not be assigned to senses themselves in his theory. He had to introduce a different faculty of mind just allocated for making judgment: the faculty of judgment. However, judgments that underlie perception are unnoticed, unconscious, rapid and habitual. In other words, they seem to be different than ordinary judgments such as the visual theorist's careful and deliberate judgment about perception requiring unnoticed judgments. Thus, at first glance, it seems that Alhazen needs a different faculty of judgment for ordinary deliberate judgments. Contrary to this seemingly necessary requirement, Alhazen argued that it is the same faculty of judgment that conducts both unnoticed, rapid perceptual judgments and ordinary deliberate judgments. In fact, Alhazen claimed that unnoticed,

rapid perceptual judgments are similar to syllogisms that we use in ordinary judgements. The minor premise of such syllogisms comes from the passive apprehension of stimuli, and the major premise is a result of previous learning. Using these two premises, the faculty of judgment concludes the required content of perception. Since such perceptual judgments are rapid and habitual they go unnoticed. Claiming that perceptual judgments are syllogism-like, however, presents another difficulty for Alhazen. Syllogisms are expressed by linguistic entities whereas unnoticed perceptual judgements seem to be non-linguistic. Such a difference, if it truly exists, could be enough to reject Alhazen's claim. Alhazen's solution of this problem is that even ordinary syllogisms are non-linguistic despite the fact that we express them in natural languages. He argued that in most of our ordinary syllogisms we reach a conclusion very rapidly without consciously constructing every step of the logical inference. One of the examples that he used is the following. When we hear someone exclaim 'How effective this sword is' we quickly conclude that the sword is sharp without carefully constructing any syllogisms. We hear the exclamation which gives us a particular claim, and we remember the universal claim 'Every effective sword is sharp', and then immediately we reach the conclusion. This inference is done without using words or ordering the premises. Despite the fact it is not as unnoticed as perceptual judgments, it has a substantial similarity to quick and rapid perceptual judgements. Hence, he concluded, assigning the same faculty both for unnoticed and ordinary deliberate judgments is not problematic.

In short, in Alhazen's theory, quick and unnoticed perceptual judgments are made by the same faculty that makes our ordinary judgments which are expressed by syllogisms. Neither ordinary syllogisms nor rapid perceptual ones necessarily include linguistic entities. Most of our ordinary syllogisms are also carried out in a quick and rapid manner as unnoticed perceptual judgments. Several features of Alhazen's theory are open to debate. Whether one really needs to assume the existence of the faculty of judgment or not; if the faculty of judgment exists, what are the characteristics of that faculty and so on. For the purposes of this dissertation, I do not have to delve into these questions. The only relevant point for my purposes is the characteristic of unnoticed perceptual judgments. Alhazen's claim about

such judgments being syllogism-like is not acceptable for the theory that I develop since it is an inherently probabilistic account. Unnoticed and unconscious perceptual inferences need to be inductive inferences. This is the main point where my analysis of unconscious inference diverges from Alhazen's theory.

2.2. Descartes. Descartes described his theory of vision his *Optics* [Descartes 1984-85b]. He considered visual perception as the paradigm of all other sense modalities. Descartes' theory of vision is a continuation of Alhazen's theory. Descartes also appealed to unnoticed and unconscious judgments as the underlying mechanism of visual perception especially of size, shape and distance perception. Since he considered visual perception to be the paradigmatic example of all sense modalities, his ideas for visual perception are true of other sense modalities as well. One passage in his *Optics* clearly shows his appeal to the judgmental character of the human sensory system.

> [the size of an object is identified] by the knowledge or opinion we have of their distance, compared with the size of the images they imprint on the back of the eye-and not simply by the size of these images

The characteristics of the judgmental process involved in perceptions are specified in Descartes' sixth set of *Replies to Objections to the Meditations* [Descartes 1984-85a]. Such judgments acquired through repetition; they, thus, are habitual and rapid. Despite this fact, says Descartes, these judgments are made exactly in the same manner of reflexive and deliberate judgments. Like Alhazen, Descartes also posited one faculty of mind for both types of judgments: the faculty of intellect. Sensory stimuli provide us with a passive reception of images about shape, color and distance. Through the developmental process, from infancy, we learn rules about the relationship between these images and the actual values of shape, color and distance. The unnoticed perceptual judgments are based on these two streams of information. In a sense, like Alhazen, Descartes claims that the minor premise of the perceptual inference is provided by passive apprehension of the stimulus and the major premise is provided by our developmental learning process. Through repetition these rules

and the judgmental process become so habitual to the point that we no longer recognize our perceptions as being judgementally based.

Positing a faculty of intellect, or a faculty of judgment as in the case of Alhazen, sounds circular to the ears of philosophers. As the argument goes, we try to explain how make judgments by appealing to a faculty of judgment! However, as Hatfield claims in his survey article, this is not exactly what theoreticians like Alhazen and Descartes were trying to do. Their goal was to identify the primitives by which one could explain perceptual processes. For this purpose, they focused on the interaction between the capacity of rational inference and the passive reception of sensory information. The dialectic between these two explains how sensations are transformed to lead to perceptions of shape, color and distance. They did not posit the faculty of intellect in order to explain how the mind reasons; rather they posited such a faculty for identifying the constituents of perception. Thus, the circularity objection does not hold any ground.

The other important issue about Descartes' faculty of intellect is that since the same faculty is functional both in unnoticed perceptual judgments and ordinary reflective judgments, it is legitimate to ask why the rapid and habitual perceptual judgments are not subject to modification. The reflective ordinary judgments, made by the faculty of intellect, can be modified upon receiving new information. On the other hand, the rapid and habitual perceptual judgments cannot be; visual illusions are good examples of this fact. We still perceive the stick submerged in water as bent despite the fact we know that it is a straight stick. Descartes claimed that the rapid and habitual perceptual judgements are not open to conscious revision because they are so ingrained in our perceptual system. Despite the fact that he did not use frequency based terminology, it would not be wrong to say the following: the number of times that our perception of shape is faithful to the external stimulus significantly outweighs the number of unfaithful instances. Thus, the rapid and habitual judgments are impervious to knowledge. This way of phrasing Descartes' explanation fits well the framework of my theory, since probabilities at the diachronic level are calculated through the relative frequencies which are the result of the organism's past experiences.

As in the case of Alhazen's faculty of judgment, I do not have to endorse Descartes' faculty of judgment for utilizing the idea of perception as unconscious inference in the theory of mental content that I develop in this dissertation. The main point that I borrow from Descartes' theory of perception is that the rapid and habitual perceptual judgments are impervious to knowledge and they are not recognized as judgments because of being very habitual.

2.3. Helmholtz. As mentioned above, Helmholtz provided the modern formulation of the idea of perception as unconscious inference. The primary statement of his formulation is in his *Handbuch der physiologischen Optik* [Helmholtz 1971, 1867]. He applied the idea of perception as unconscious of inference to his analysis of space perception in his *The Facts of Perception*. His original motivation for invoking this idea was to give an empirical analysis of space perception and reject the Kantian legacy about innateness of space as a form of intuition. I discuss this issue in detail in Chapter 3. Then, he also used this idea for explaining visual illusions.

The way Helmholtz characterized the inferential process that underlies perception significantly differs from Alhazen's and Descartes' characterizations. He combined 'the associative and inferential accounts by giving an associational account of inference' [Hatfield 2002, p.124]. He argued that the major premise of the inference is acquired inductively through association, and he compared the inductive process with hypothesis testing in science. His analysis of visual sensation exemplifies his associational account. In his analysis, every activated nerve fiber varies in three dimensions: hue, intensity and local sign. Local signs are qualitative markers of nerve fibers. They don't have any inherent spatial meaning. However, coordination of these nerve fiber activations with bodily movements and sense of touch, local signs acquire spatial meaning. At the end, the observer learns spatial meanings of local signs through association. This gives to the observer the major premise that is required in the unconscious perceptual inferential process. In Helmholtz's own words,

> while in these cases [referring to learning spatial meaning of local signs] no actual conscious inference is present, yet the essential and original office

> of an inference has been performed ... The inference is achieved simply, of course, by the unconscious processes of the association of ideas going on in the dark background of our memory. (3:24)

Helmholtz's characterization of the idea of perception as unconscious inference is the best candidate for the needs of the theory that I develop in this dissertation. The inductive and associationist character perfectly fits the probabilistic approach, and his emphasis on learning and memory is in line with the diachronic level that is formulated above; as we will see in the case of Irvin Rock not every theory that uses the idea of perception as unconscious inference emphasizes the importance of learning and memory.

2.4. Irvin Rock & Fodor. Rock's theory of cognition is the most explicit use of the idea of perception as unconscious inference in current literature [Rock 1975, 1983]. Fodor also invokes the idea in his language of thought hypothesis [Fodor 1975, 1983]. Rock's and Fodor's theories have a lot in common, therefore I will present them together.

In his *Logic of Perception*, Rock discusses several different types of cognitive operations that are functional in unnoticed perception. Hatfield in his historical survey of the idea of perception as unconscious inference categorizes Rock's types of cognitive operations into four:

> (1) unconscious description, in the case of form perception; (2) problem solving and inference to the best explanation, in the case of stimulus ambiguity or stimulus features that would yield unexplained coincidences if interpreted literally; (3) relational determination of percepts, such as those involved in perceiving lightness and relational motion through the interpretation of relational stimulus information in accordance with certain assumptions; and (4) deductive inference from a universal major premise and an unconsciously given minor premise, used to explain constancies [Hatfield 2002, p.125]

According to Rock, the rules that govern perceptual inference involved in these cognitive operations may be either learned or innate. Including innateness in the picture implies that Rock rejects Helmholtz's emphasis on learning.

In Rock's formulation, all four types of perceptual inference operate linguistically and follow the rules of predicate logic. The linguistic medium that carries out these cognitive operations is an innate language of thought. Fodor also posits an innate language of thought as a result of his view of the mind as a general-purpose digital computer. The digital computer metaphor suggests the idea that perception follows the rules of logic in Fodor's theory. Moreover, both Rock and Fodor argue that the unnoticed perceptual judgments and the ordinary reflective judgments should be made with different modules of the mind because perceptual judgments are impervious to knowledge. As a result, Fodor develops the idea of encapsulated modules and Rock formulates the idea of insulated modules. In short, Rock's and Fodor's theories have three main assumptions in common.

- an innate language of thought
- perception follows the rules of logic
- encapsulated or insulated modules

Because of the first and the third assumptions, Rock's and Fodor's theories end up being less parsimonious than Alhazen's and Helmholtz's theories. This is the main reason the Rock-Fodor way of interpreting the idea of perception as unconscious inference is not a good candidate for the theory I develop in this dissertation. The other reason is their emphasis on the innateness of the rules by which inductive claims involved in perceptual inference. Without giving much justification, I adopt Helmholtz's emphasis on learning in this dissertation.

In conclusion, in this section I presented a brief historical survey of the theories that interpret perception as unconscious inference. Moreover, I discussed my reasons for adopting Helmholtz's theory among the others for the probabilistic theory of mental content that I develop in this dissertation. In the next section, I briefly discuss my solution for the

problem of misrepresentation, and then analyze three possible objections to the solution that I defend.

3. The Solution

As eloquently put by Fodor, in order to account for misrepresentation cases, it must be the case that 'the conditions for the truth of a symbol dissociate from the conditions whose satisfaction determine what the symbol represents' (for Fodor, mental representations are symbols)[Fodor 1992, p.42]. In other words, the dissociation between the truth conditions and the content assignment conditions (or the representation conditions) is necessary for explaining away the phenomenon of misrepresentation. The general framework of causal approaches, unfortunately, does not provide the required dissociation. Both the truth conditions and the content assignment conditions turn out to be one and the same in that framework. There have been several attempts to solve this problem within causal/informational approaches. The main attempts are Fodor's Asymmetrical Dependency, Dretske's learning versus retrieval phase distinction and the teleosemantical approach. None of these attempts successfully accounts for misrepresentation cases. I have discussed these solutions in the previous chapters. However, to discuss these solutions very briefly is useful for setting up the stage for the solution that I offer.

Fodor claims that misrepresentation cases asymmetrically depend on true representation cases. In other words, we have R_2 because of R_1.

R_1: Dogs cause DOG.

R_2: Cats cause DOG

The relation between dogs and DOG is nomically independent, but this is not true of the relation between cats in a dark night (which might be mistaken for dogs) and DOG. Given this asymmetrical dependence, the content of a mental representation is fixed by R_2 in misrepresentation cases and the truth conditions are determined by R_1. Hence, the required dissociation is satisfied. As I argued in Chapter 1, what Fodor says is nothing but question begging. Solving the problem of misrepresentation means to give an explanation why misrepresentation cases depend on accurate representation cases but not vice versa.

Thus, Fodor's asymmetrical dependence hypothesis seems to be providing the required dissociation but lacks an independent justification for the idea of asymmetrical dependence.

Dretske in his pre-1986 works attempted to solve the problem of misrepresentation by the distinction between the learning period and retrieval period of an organism. As the reader may remember, in Dretske's notion of informational content and mental content the conditional probability of the correct external state of affairs given a mental representation is 1. Because of this relationship, it is impossible to dissociate the truth conditions from the content assignment conditions. However, he claimed that after the main learning period, the organism starts using more lenient constraints, that is to say, the conditional probability in the definition of informational content is not 1 anymore. It is true that this opens up a door for dissociating the truth conditions and the content assignment conditions. It is possible to separate these two since the latter is determined probabilistically and the former causally. Dretske's attempt relies on finding a point where the learning period stops for the organism. However, it is rather obvious that there is no such point for any organism that is of interest to us.

Probably, the best among all solution attempts is the one based on teleosemantics. The teleosemantical approach has been the dominating theory in the mental content literature more or less since 1986. Its solution to the problem of misrepresentation seems very appealing at first glance. Here is how it works. Any mental representation is a state of biological mechanism. The proper function of the mechanism is what determines the content of a mental representation. Take the DOG representation as an example. Such a representation is a state of a biological mechanism, say M. DOG is produced by the mechanism M. The proper function of M is to produce DOG when presented by dogs. This is why M is evolutionarily selected. On the other hand, under certain circumstances the mechanism M may produce DOG when there is a cat in the visual field of the organism. The truth conditions for DOG are determined by the entity that triggers the DOG state of the mechanism M. Hence, in such a situation the truth conditions of the mental representation DOG are dissociated from the content assignment conditions. Once again, the content is determined by

the proper function of the mechanism M and the truth conditions are determined by the entity that causes M to produce DOG state.

The main problem with the teleosemantical solution is the notion of proper functions. Proper functions are normative and in order to provide a naturalistic account of mental content, one needs to clarify how proper functions get their normative character. Until this is done, the teleosemantical approach is just a step in the right direction but not a complete solution. An analogy from linguistic content is helpful here. The teleosemantical approach tries to explain mental content by referring to the proper functions of biological mechanisms. John Locke tried to explain linguistic content by referring to the ideas in the mind. To put it very simply, according to Locke, a word has its linguistic content via the idea that it refers to. What the teleosemantical approach does is pretty analogous to John Locke's explanation of linguistic content. On the other hand, it is almost commonly accepted that what John Locke says about linguistic content is not wrong, but an incomplete explanation. Unless a proper analysis of the notion of ideas is given, Locke's explanation of linguistic content will be incomplete. I think that a similar problem is true of the teleosemantical approach as well. How do we acquire proper functions with their normative properties is the main question that the teleosemantical approach needs to answer. The conclusion that I draw from this analogy is that the teleosemantical approach is a step in the right direction, but not a solution.

The solution that I offer for dissociating the truth conditions of a mental representation from the content assignment conditions relies on a very simple idea: the distinction between the instantaneous interaction with the external world and the inference made based on the previous knowledge of the organism. In other words, the required dissociation is guaranteed by exploiting the distinction between the organism's short term and long term interactions with the external world. The truth conditions are fixed by the instantaneous interaction with the external world. The content assignment conditions, on the other hand, are fixed by the unconscious inference made on the basis of the instantaneous interaction and the previous history of the organism. Now let me state this solution in detail.

The overall goal of a theory of mental content is to give a naturalistic account of mental representation. Then, the natural starting point is the causal interaction between the external world and the organism that has mental representations. There are several other reasons for starting with the causal interaction between the organism and the world. I discussed them in the first chapter. The story starts with an external state of affairs, say S, causing a neurobiological / mental entity, say R, in the mental realm of the organism. Because of the causal connection, R has a special relation with S. As we discussed in earlier chapters, this relation falls under the category of indicative relations.

'S causes R' and 'R indicates S'.

As Cummins & Poirier [2004] say indication is just a semantic-sounding word for detection. That is to say, by R the organism is able to detect the presence of S. However, this detection cannot provide enough grounds for representation relation. This is precisely because of the special characteristics of indication relation. Cummins and Poirier provide a careful analysis of these characteristics in their *Representation and Indication* article. They identify three distinguishing features of indication relation: transitivity, source dependence and arbitrariness. For my purposes, transitivity and source dependence are the essential ones. Thus, I shall explain only those two by quoting Cummins and Poirier.

> Indication is transitive, representation is not. If S3 indicates S2, and S2 indicates S1, then S3 indicates S1. Aim a photosensitive cell at the oil pressure indicator light in your car. Attach this to a relay that activates an audio device that plays a recording of the sentence, "Your oil pressure is low." If the light going on indicates low oil pressure, so does the recording. Indeed, there is already a chain of this sort connecting the pressure sensor and the light. Representation, on the other hand, is not transitive. A representation of the pixel structure of a digitized picture of my aunt Tilly is not a representation of my aunt Tilly's visual appearance, though, of course, it is possible to recover the later from the former. To anticipate some terminology we will use later, a representation of the pixel structure

> is an encoding of my aunt Tilly's visual appearance. [Cummins & Poirier 2004]

For source dependence they say the following.

> Indicators are source dependent in a way that representations are not. The cells studied by Hubel and Weisel all generate the same signal when they detect a target. You cannot tell, by looking at the signal itself (the spike train), what has been detected. You have to know which cells generated the signal. This follows from the arbitrariness of indicator signals, and is therefore a general feature of indication: the meaning is all in who shouts, not in what is shouted. [Cummins & Poirier 2004]

Once again, the indication relation is an obvious starting point for a naturalistic theory of content. The synchronic level, which I discussed in Chapter 4 and in the first section of this chapter, aims to explain this relation. At any given moment of perception, the organism acquires some information from the external world. Following Irvin Rock's terminology, we can call this the stimulus-based information. The relationship between the external state of affairs, S, and the mental entity, R, is one of indication. This relation is transitive as Cummins and Poirier point out. The transitivity at this level corresponds to Dretske's Xerox principle. Hence, the conditional probability used in the definition of synchronic informational content is 1. So far so good, but this cannot be the whole story. If we stopped at this level, like Dretske did in his 1981 framework, we would not be able to provide the required dissociation between the truth conditions and the content assignment conditions. Here is how

'S causes R' and 'R indicates S'.

The entity that R truthfully applies to is S itself. On the other hand, the content of R is determined by S. Hence, the truth conditions and the content assignment conditions are one and the same. If Fodor's dissociation requirement is necessary for misrepresentation cases, then this indication relation is not sufficient. This is exactly what Cummins and

Poirier call the source dependency of indication relation. Indicative signs cannot be completely separated from their source. On the other hand, representations must be portable in a way suitable for structural processing. This is why I introduced the diachronic level. This the level where the indicative signals from the synchronic level are processed through the memory mechanisms of the organism. As a result, they become source independent. Another way of seeing this source independence is the conditional probability used in the definition of informational content at the diachronic level. It is less than 1. Let me symbolize the relevant knowledge stored in the memory mechanism of the organism by K. Then the process could be summarized as follows.

The Synchronic Level: R indicates S (the stimulus-based information)

The Diachronic Level: T is inferred from R and K. (unconscious inference)

On the one hand, the content assignment conditions are determined by R and K. In other words, T's content comes from the inference based on R and K. On the other hand, the truth conditions are determined by S (the external entity that caused R). That is to say, the entity that T truthfully applies to is S. Since 'S' is different from 'R and K', the dissociation requirement is fulfilled. This is the gist of the solution that I offer for the problem of misrepresentation. However, more explanation about the process on the diachronic level is needed.

The causal interaction between the world and the organism starts at the time that the organism comes into being. The organism acquires its mental entities through its causal connection with the world throughout its ontogenetic evolution. Concepts, ideas, rules are formed via a basic statistical generalization over the instances that the organism encounters throughout its life. The frequency of each category and concept depends on the history of the organism. This frequency is constantly updated with new experiences. And such frequencies are the basis of the probabilities that are assigned to each category at the diachronic level. As an example, assume that at a given moment of its history, the organism faces an external state of affairs, S. The organism acquires stimulus-based information from S. We symbolized this stimulus-based information by R. R could be consistent with several categories and/or concepts stored in the memory system of the organism. The organism selects the relevant

alternatives from its huge repository of concepts and categories. I formalized the notion of relevance with the idea of minimal inconsistency in Chapter 4. This mathematical notion gives us the power of identifying the relevant categories. In our example, all the members of the minimally inconsistent set are content assignment candidates for R. The minimally inconsistent set is identified through R, which is the stimulus-based information. Say, for the sake of simplicity, this set includes three members T, Q and P. The probabilities of each of these candidates given R are determined by the past experiences of the organism. And the organism selects the one with the highest probability. This is what I called the ordinal ranking approach in Chapter 4. The selection of the one with the highest probability is an ampliative inference, and the inference has two premises: the stimulus-based information and the categories stored in the memory system of the organism. In our simple example, the former corresponds to R and the latter corresponds to the set that includes T, Q and P. The characteristics of the ampliative inference at the diachronic level are analyzed in the first section of this chapter. It is an unconscious inference, and the organism makes such inferences at every moment of perceptual interaction with the world.

As mentioned above, such a two-level analysis of mental representation fulfills Fodor's requirement of dissociating the truth conditions of a mental representation from the content assignment conditions. Once again, the truth conditions are determined by the stimulus-based information and the content assignment conditions are determined by the inference that is made by two premises: the stimulus-based information and the relevant categories stored in the memory of the organism. On the other hand, since the inference is ampliative, then there is a possibility of making mistakes in a substantial manner, that is to say, in a way that gives us the source independency requirement for representation relations.

Let me try to visualize the solution that I offer for the problem of misrepresentation. The synchronic level is purely naturalistic and falls under the category of indication relations. This implies that the signal is source dependent, as Cummins & Poirier [2004] claim. These indicative instances stem from nomic regularities. This is what Grice calls natural meaning instances. When the information from this level is subjected to an inductive inference together with stored information, then the representation relation (and thereby the

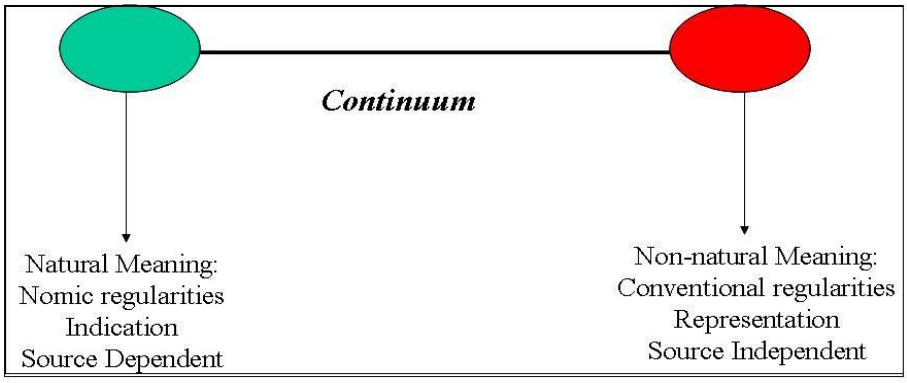

FIGURE 1. The Continuum between Indication and Representation

misrepresentation relation) emerges. The representation relation is not source dependent anymore because of the effect of the stored information in the inferential process. There is a continuum between the indication and the representation instances. The extreme cases of representation relation stem from conventional regularities where the amount of stored information required for the inference process is the highest. This is what Dretske and Grice call non-natural meaning instances. Dretske and Grice, as mentioned in Chapter 4, considered these two categories in a dualistic manner. However, as I argued, this is the wrong interpretation. The relation between these two extremes (purely indicative and purely representational) is a continuum. Figure 1 depicts this relation.

The way one identifies the location of a given mechanism on the continuum line needs to be specified. Since the ultimate difference between the synchronic level and the diachronic level is the unconscious inference, the amount of the memory knowledge involved in the unconscious inference should be the criterion by which one identifies the location of a given sensory mechanism on the continuum of the representation relation. This is why the title of this dissertation is 'Error Comes With Imagination'. Figure 2 summarizes this idea. In this framework, any organism with a sensory mechanism that uses past experiences in an inferential way qualifies for representation and misrepresentation, but in different degrees. Thus, our marine bacterium, if it turns out that the magneto taxis mechanism essentially

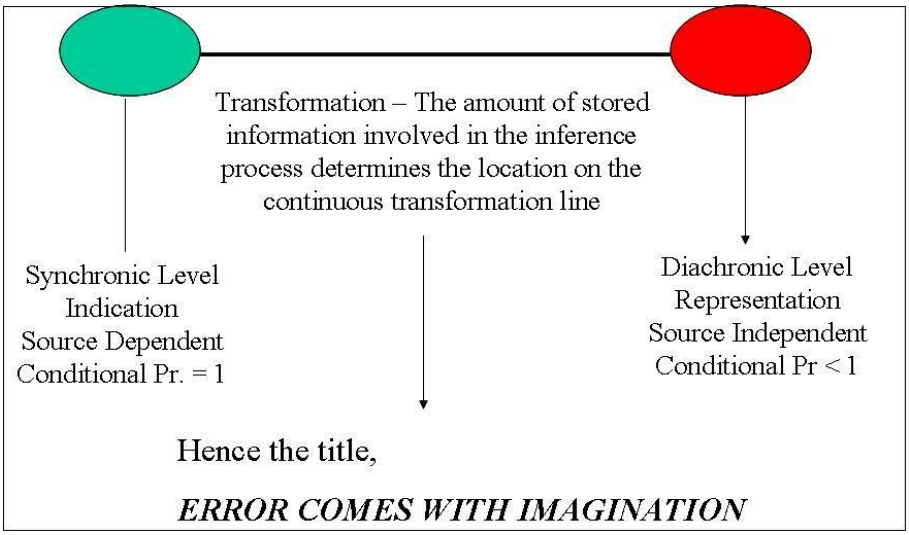

FIGURE 2. Error Comes with Imagination

uses its past experiences in an inferential way, may have the capability of representation and misrepresentation.

4. Three Objections

4.1. Objection 1:Disjunctive Categories. One objection that could be raised against the solution that I offer is about disjunctive categories. Once we have set categories stored in the memory such as CAT and DOG in our toy example, then the possibility of misrepresentation is easily accounted for. However, one could question the type categories that the organism has. Instead of CAT and DOG categories stored as mental representation categories, why do not we have disjunctive categories like CAT OR DOG? If we started with disjunctive categories, then the disjunction problem which is the sister problem of misrepresentation would arise. This line of objection needs to be refuted in order to prove that the probabilistic theory that I offer genuinely solves the problem of misrepresentation. Since my probabilistic theory analyzes perception in two levels, I have to show that categories are not disjunctive at each level.

The synchronic level is where the organism acquires the stimulus-based information from the external world via causal links. Nature does not present itself in disjunctive

categories. In other words, there is no disjunctive entity in the world that can present a causal unity. The world presents itself to organisms in the form of causally efficacious units, and disjunctive categories do not qualify as causally efficacious units. The shape of an object (say a dog) can causally influence our sensory mechanisms or the color of an object can do the same thing, but there is no causal unity of 'dogs and cats' that can causally effect our sensory mechanisms. Thus, it is not possible to form disjunctive categories at the synchronic level of our perceptual processes.

It seems that there is no room for disjunctive categories at the synchronic level because of the structure of the external world. However, it is still possible to have disjunctive categories at the diachronic level. That is to say, despite the fact that the stimulus based information does not come in disjunctive forms, it is still possible that the organism forms disjunctive categories out of the stimulus-based non-disjunctive information. There are two reasons that speak against this possibility. The first one is related to the notion of minimal consistency which is the basis of the unconscious inference of the diachronic level. The second one stems from the information maximizing characteristic of the inference process.

The organism 'selects' the proper content for stimulus-based information among the relevant alternatives, and the relevant alternatives are listed as the members of a minimally inconsistent set. Thus, in order for disjunctive categories to be legitimate alternatives they need to be members of a minimally inconsistent set, otherwise they would be not eligible for selection for content assignment. Let's take a simple example where a disjunctive category, say A or B (A $\bigvee$ B), is a relevant alternative.That is to say, (A $\bigvee$ B) is a member of the minimally inconsistent set that is functional at the diachronic level. If, however, the disjunctive category is relevant, then each of its disjuncts should also be relevant for content assignment. Then, both A and B are also members of the minimally inconsistent set. So far, we have three relevant alternatives. As explained in the previous chapters, to satisfy the minimal consistency requires using the negations of the alternatives. Thus, we have the following set given the three alternatives.

The Set $= \{\sim A, \sim B, \sim (A \vee B)\}$

In order for this set to be inconsistent, the probability of the conjunction of all of its members must be zero. Thus,

$$\mathrm{P}(\sim A \wedge \sim B \wedge \sim (A \vee B)) = 0$$

However, from the basics of Propositional Logic, we know that $(\sim A \wedge \sim B \wedge \sim (A \vee B))$ is logically equivalent to $(\sim A \wedge \sim B)$. Therefore, their probability values should be identical.

$$\mathrm{P}(\sim A \wedge \sim B) = 0$$

This implies that the set that includes $\sim A$ and $\sim B$ is also inconsistent, but the set that includes $\sim A$ and $\sim B$ is a proper subset of the original set. Hence, the original set that includes the disjunction of A and B as a relevant alternative is not minimally inconsistent. This is true for any set that has such a disjunction of some other members of the set. If the original assumption (if a disjunctive category is relevant for content assignment, then so is any of its disjuncts) is correct, then adding a disjunctive category of its members to a set violates the minimal inconsistency requirement. Thus,since adding a disjunctive category violates the minimal inconsistency requirement, it is safe to say that disjunctive categories do not appear at the diachronic level.

The second reason why disjunctive categories are not legitimate alternatives for content assignment comes from the idea of information maximization. The main motivation behind selecting the alternative with the highest probability among relevant alternatives at the diachronic level is to maximize the amount of mutual information (or minimize the Kullbeck-Leibler divergence). Non-disjunctive categories, compared to disjunctive ones, provide more information to the organism. Austen Clark elegantly proves this fact in his *Mice, Shrews and Misrepresentation* [Clark 1993]. Let me briefly present Clark's simple example by which he proves this fact.

In Clark's simple example, the world consists of three non-overlapping object types: S, T and M. S and T each has 50 instances and M has 100 instances. Given this structure of the world, Clark imagines two organisms: one with disjunctive categories and one without disjunctive categories. Organism 1 has only two categories in its 'mental realm': 'S or T' and M. When this organism is presented with an instance of S or with an instance of T,

then the disjunctive category of 'S or T' is triggered. The amount of information that is generated by any of these mental categories is exactly one bit (in terms of the amount of information, the case of Organism 1 is identical to a coin toss). The second organism also has two mental categories, but no disjunctive categories. The mental categories that Organism 2 has are S and M. Since there is a third object type in the world, namely T, one needs to introduce noise into the picture as well for T instances that might trigger the mental category of S. How many of T instances should be considered as noise is up to us. When none of the T instances is considered as noise, then Organism 2 is identical to Organism 1. However, as the number of T instances that are considered as noise increases, so does the amount of information that Organism 2 acquires from the world. After the amount of information that Organism 2 acquired is calculated, the effect of noise could be excluded mathematically. In both cases, with keeping the noise factor in the amount of information and with excluding it after the fact, Organism 2 fares better than Organism 1 in terms of the amount of the information that is acquired from the world. Figure 3 presents the mathematical results for both Organism 1 and Organism 2 graphically.

Thus, non-disjunctive categories produce more information than disjunctive categories. If our mental system is geared towards information maximization then it is only normal for our mental systems to have non-disjunctive categories.

In conclusion, disjunctive categories are not legitimate alternatives at the diachronic level. I presented two reasons for this claim: the minimal inconsistency requirement and Austen Clark's proof about information maximization. Hence, the disjunction problem which is the sister problem of the problem of misrepresentation does not present a justified objection to the probabilistic theory that I offer.

4.2. Objection 2:Doxastic States. In the probabilistic theory that I offer, the knowledge that is stored in the memory system of the organism plays an essential role at the diachronic level. The main requirement of a successful theory of mental content is to provide a naturalistic account of mental representation. However, since there is an essential

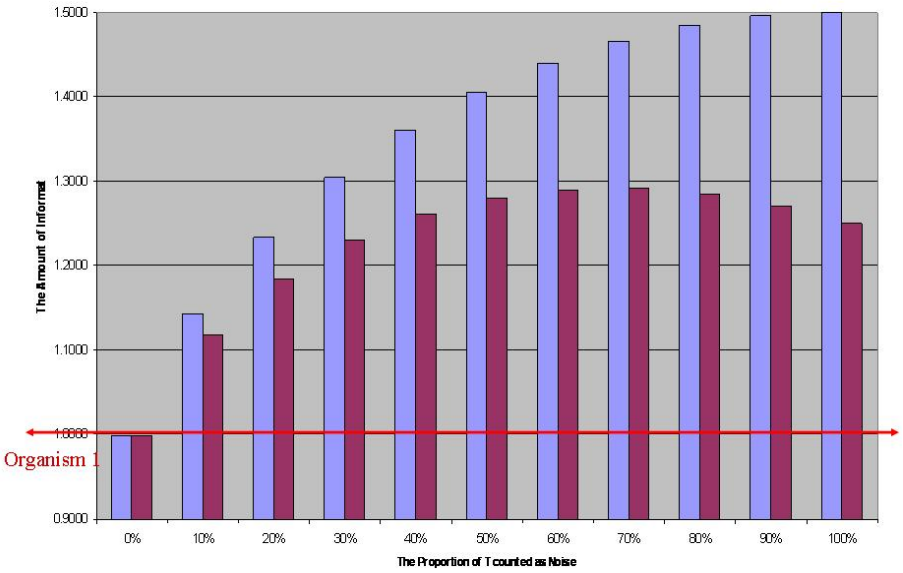

FIGURE 3. Blue-Light Bar: Organism 2 with Noise. Purple-Dark Bar: Organism 2 without Noise

reference to memory in the theory that I offer, one could question whether it is a naturalistic theory or not. In other words, one objection is that because of the essential reference to the background knowledge of the organism (i.e. to doxastic states), my theory ends up being non-naturalistic. Such an objection is pretty similar to the doxastic states objection that Cohen & Meskin [2006] raise against Dretske's theory.

This line of objection does not hold any ground within the context of my theory. Despite the fact that the diachronic level operates wholly on doxastic states, the synchronic level has no such reference. Everything that happens at the synchronic level is naturalistic, and is a result of the purely causal interaction between the organism and the world. Moreover, doxastic states that are functional at the diachronic level are formed through those interactions via learning mechanisms (this is precisely why I adopt Helmholtz's interpretation of perception as unconscious inference among other alternatives). Hence, the reference to doxastic states at the diachronic level could easily be reduced to natural causal interactions of the synchronic level. The way to do this is by a simple backwards reiteration.

4.3. Objection 3: Unprincipled Distinction. Dretske's original solution for the problem of misrepresentation relied on the distinction between the learning and the retrieval periods of the organism. He said that in the learning period, the organism was not vulnerable to making mistakes, and that fit well to assigning unity to conditional probabilities in his definition for informational content. After the learning period, he argued, the organism relaxes its constraints about conditional probabilities, and becomes vulnerable to making mistakes. Thus, misrepresentation is possible only after the learning period. This solution came under immense attack, because the distinction between the learning and the retrieval periods is unprincipled. When does the learning stop? Dretske did not have a satisfactory answer for this question. This is precisely why his distinction is unprincipled. Now, since there is an emphasis on learning in the theory that I offer one might suspect that this line of objection applies to my theory as well. I claim that this suspicion is unfounded. In the theory that I offer, there is no distinction between learning period and retrieval period. The reference to learning is constantly in play throughout the entire life span of the organism. Each moment of content assignment (or unconscious inference at the diachronic level) is stored in the memory system of the organism, and it is carried over to the next moment of content assignment. Hence, I do not need to make a distinction between the learning period vs. the retrieval period.

Another question about unprincipled distinctions might arise with respect to the distinction between the synchronic and the diachronic levels. The main variable that distinguishes these two levels is whether there is an inference or not. In other words, whether the background knowledge of the organism is functional in the content assignment or not is what distinguishes the synchronic level from the diachronic level. Since the variable that distinguishes these two levels is not arbitrary, the unprincipled distinction objection does not apply here, either.

5. The Regress Problem: An Incompleteness Claim

The solution that I offer in Section 3 is based on a two-level analysis of informational content of mental representations. In order to provide a naturalistic theory of mental

content, the two-level analysis starts with indicative relations of the synchronic level. All other entities that are involved in the analysis are supposed to be derived (or formed) through those indicative relations. For example, concepts and ideas that provide us the second premise of the inference at the diachronic level are formed through the indicative instances of the synchronic level. This is a bold empiricist claim. Given this claim, the theory that I offer will remain incomplete until I explain how indicative instances lead to concepts and categories such as CAT, DOG, etc. I think that such an explanation requires providing an ontological framework. This by itself is a dissertation topic. Unfortunately, I am not able to do that within the scope of this project. The only thing that I can do is to clarify the issue and suggest some clues about an ontological framework that has the potential of providing the required explanation.

As mentioned above, every category that an organism uses is formed through the indicator signals of the synchronic level. These indicator signals carry information about the external world. Their content is determined through the entity that causes them. This is why the conditional probability at this level is one, and this is why misrepresentation is not possible at this level. What are the external entities that can trigger indicator mechanisms of an organism? Since the theory that I offer is a strictly causal theory, these entities must be causally efficacious entities. One possible answer is basic properties of the ontological inventory of the world. However, if this is the case, then I have to give an account of concepts and ideas that are used at the diachronic level in terms of these basic properties. For the sake of simplicity, imagine a simple world where there are five basic properties. Further assume that, these properties are the following: fury, has tail, four legged, barks and meows. Obviously, this simple world has two objects in it: cats and dogs. Dogs provide information about the first four properties, and cats about the first three and the fifth one. When an organism in this world encounters some instances of cats and dogs, it faces with the problem of categorizing over these instances. Why should the organism choose the categories of CAT and DOG instead of FURY, HAS TAIL etc? Another similar question, why should the organism choose CAT and DOG instead of 'CAT OR DOG'? A different way of stating the problem is the following: once we have the categories like CAT, DOG, HORSE etc., the

theory that I construct in this dissertation solves the problem of misrepresentation. Then, the theory should provide an account for reducing these categories to indicator relations of the synchronic level. Borrowing Prof. Frederick Schmitt's coinage, I call this the regress problem. Unless the regress problem is solved, the theory that I offer will be incomplete.

One way of dealing with the regress problem could be to provide an ontological framework of the world that forces the organism to form the proper types of categories. A simple example is useful here. For the sake of simplicity, assume that the ontological inventory of the world consists of fundamental properties and bundles of these properties at specific locations. Moreover, assume that the informational content that an organism acquires from the world is consistent with the ontological structure of the world. Then it might be reasonable to assume that the organism forms categories like FURY, HAS TAIL as well as CAT, DOG, HORSE etc. If this is right, then finding the right ontological framework is the way to solve the regress problem. Let me state a disclaimer before I proceed with a suggestion regarding the right ontological framework. I don't claim that an ontological framework that has the two aforementioned features will definitely solve the regress problem. I simply don't know. One needs to construct such a framework and test it within the context of the theory that I construct in this dissertation. Unfortunately, I cannot do it within the scope of this dissertation. This story needs to be told at another time, but it definitely needs to be told. The only point that I make in this section is that the regress problem has to be dealt with and a proper ontology might provide the required explanatory tools. Now, it is time to say a thing or two about the ontological framework that I have in mind.

For the ontological framework, Denkel's particularistic view seems to be promising [Denkel 1996]. Denkel, following Locke's footprints, considers properties as the basic analytic units of the world. As is well known, the main problem of the Lockean ontology is that its principle of individuation, which is based on the notion of substrata, is not empirically accessible. Denkel rejects this principle, and retains other features of the Lockean ontology. He defines objecthood as a co-presence of properties at a specific location. This formulation is able to provide a satisfactory principle of individuation. In Denkel's ontology, only particular properties exist in the physical world, and when some properties are co-present

at a location, they become an object. Moreover, some objects resemble each other in terms of their constituent properties. These objects form a resemblance class, and such a class is the basis of what we call a natural kind.

Causality as objecthood is also analyzed in terms of particular properties. Any change in the world is a property occurrence: one property gives rise to another one. Causal instances, since they include change, are also property occurrences. An instance of a causal relation is treated as a complex structural property where the relata of it are properties that reside in a co-presence. This framework allows us to analyze any instance of causation in terms of its constituent properties, i.e. the relata. Although objects, as co-presence of properties, are necessary for an instance of a causal relation, they are not the basic analytic units. The causal interaction between objects in the external world and the human mind is also analyzed in terms of properties, or rather as a property occurrence.

The hope is that Denkel's ontological framework together with the assumption that categories of the mind conform to the ontological structure of the world could solve the regress problem. Whether or not this is the case has yet to be decided ...

CHAPTER 7

Conclusion

In this dissertation, I developed a probabilistic theory of mental content. I claim that the theory that I offer solves the problem of misrepresentation which has been a long standing problem in the contemporary mental content literature. The theory that I develop falls under the category of causal/informational theories. The notion of information plays a crucial role in the theory. The basic idea is that any kind of mental entity is formed through an information-maximizing process. The notion of information that I use is from Shannon's mathematical theory of communication. Shannon's notion of information is philosophically neutral and his theory is a powerful formalism of the notion. Dretske attempted at utilizing Shannon's notion for solving philosophical problems related to mental content and mental representation in his 1981 framework. However, as I discuss in the early chapters, his framework does not provide any room for misrepresentation cases. This problem arises because of his definition of informational content. In that definition, he assigns 1 to the conditional probability of the relevant state of affairs given a mental representation. In Chapter 4, I analyze his arguments for doing so, and argue against them. This provides enough grounds for developing a probabilistic theory of mental content where conditional probabilities can be less than 1. Besides these, the theory that I develop has four other fundamental features.

- There could be two different methodologies for studying the human mind and mental content. One is in line with the methodologies of other scientific disciplines where the entity in question is analyzed from an observer's perspective. The other one is analyzing the entity in question form the perspective of the organism that forms and uses that entity. The second methodology is called the animal's perspective. I claim that the best way of understanding the notion of mental

representation, both philosophically and empirically, requires using the animal's perspective. In Chapter 3, I provide arguments for this claim and give a historical overview of this methodology.

- The main challenge for a naturalistic theory of mental content is to give an account of how normative representation relations arise from descriptive indication relations. These two types of relations are considered as two mutually exclusive categories, at least within causal/informational theories of mental content. Two examples of this general tendency are Dretske and Grice. In the theory that I develop, they are not mutually exclusive; rather there is a continuum between indication and representation relations. Such a continuum approach makes room for intermediary categories which are a mix of indication and representation relations in varying degrees. I discuss this continuum idea in Chapters 4 and 6.
- Any kind of perceptual interaction with the external world is similar to an ampliative inference. This inference is made based on two premises: the stimulus-based information and the relevant information stored in the memory mechanism of the organism. The former is acquired through the instantaneous interaction with the external world at the time of perception. The level that provides the first premise is called the synchronic level. The latter premise is supplied by the previous history of the organism, and the level that supplies this information is called the diachronic level. The inference that is made on these two premises is an unconscious inference. There are several different interpretations of the idea of perception as unconscious inference. Among these different interpretations, Hermann Helmholtz's interpretation is more suitable for the purposes of a theory of mental content. In Chapter 6, I survey different interpretations of this idea and provide reasons for choosing Helmholtz's interpretation.
- As mentioned, the unconscious inference has two premises: the stimulus-based information and the information stored in the memory mechanism of the organism. The latter must be selected from a big repertoire of previously stored information. The ones that are relevant to the former must be selected. To find the relevant

ones turns out to be a difficult task. I use a set theoretical notion for identifying the relevant alternatives: minimal inconsistency. Keith Lehrer used that notion for providing grounds for scientific explanation in 1971. I borrow his usage and incorporate into the theory that I develop in this dissertation.

Given these features, I solve the problem of misrepresentation. As mentioned in the previous chapter, Fodor says that to solve the problem of misrepresentation requires dissociating the truth conditions of a mental representation (the conditions that determine the entities that a mental representation truthfully applies to) from the conditions that determine the content of the mental representation. My theory provides the required dissociation by the distinction between the stimulus-based information and the previously stored information. In other words, the truth conditions of a mental representation are fixed by the stimulus-based information and the content assignment conditions are determined by the inference based on the stimulus-based information and the relevant previously stored information. This is a very simple move; it really is. It may even sound too simplistic to philosophically sophisticated ears. Is it really too simplistic or is it just the simple truth? A difficult question ... May be Wittgenstein was right when he said 'Why is philosophy so complicated? It ought to be entirely simple ... The complexity of philosophy is not a complexity of its subject matter, but of our knotted understanding.'

CHAPTER 8

Appendices

1. Transitivity and Conditional Probabilities

Conditional probabilities are not transitive, that is to say there is no α, where $1 > \alpha > 0$, such that $Pr(A|B) \geq \alpha$ and $Pr(B|C) \geq \alpha$ necessitate that $Pr(A|C) \geq \alpha$. Figure 1 below takes 0.9 as a possible candidate for α and shows that the property of transitivity does not hold. This result generalizes for any value smaller than 1. Figure 4 shows that the property of transitivity holds when conditional probabilities are one. As a result, Dretske concludes, since the Xerox principle requires transitivity, conditional probabilities must be one.

Conditional probabilities are defined in terms of sets. Pr (A | B) is equal to Pr (A intersection B) / Pr (B), i.e. Pr (A $\bigcap$ B) / Pr (B) which is equal to n (A $\bigcap$ B) / n (B) where n(B) is equal to the number of the elements in B. As a result of this set theoretic definition, three sets are depicted in the diagrams. The numbers in each region represent the number of elements that belong to the set that corresponds to the given region.

Pr (A | B) = n (A$\bigcap$B) / n (B) = 90 / (90 + 1 + 9) = 90 /100 = 0.9

Pr (B | C) = n (B$\bigcap$C) / n (C) = 9 / (9 + 1) = 9 /10 = 0.9

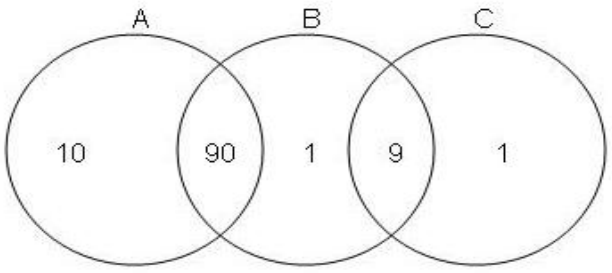

FIGURE 1. Conditional Probabilities are 0.9

$$\Pr\ (\mathrm{A} \mid \mathrm{C}) = \mathrm{n}\ (\mathrm{A} \bigcap \mathrm{C})\ /\ \mathrm{n}\ (\mathrm{C}) = 0\ /\ (9 + 1) = 0$$

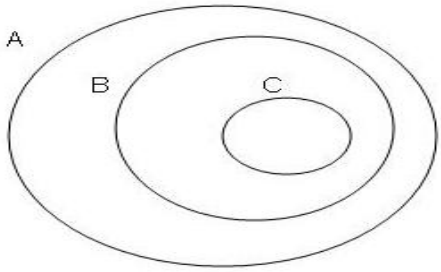

FIGURE 2. Conditional Probabilities are one

This shows that even when Pr (A | B) and Pr (B | C) are 0.9, Pr (A | C) could be as low as zero.

For the case of assigning unity to conditional probabilities, assume that both Pr (A | B) and Pr (B | C) are one.

Pr (A | B) = n (A⋂B) / n (B) = 1, hence n (A⋂B) = n (B). This implies that B is a subset of A.

Pr (B | C) = n (B⋂C) / n (C) = 1, hence n (B⋂C) = n (C). This implies that C is a subset of B. These two claims give us Figure 2. Since C is a subset of B and B is a subset of A, C has to be a subset of A. This implies that n (A⋂C) = n (C). Thus Pr (A | C) = 1. Two assumptions, Pr (A | B) = 1 and Pr (B | C) = 1, necessitate that Pr (A | C) = 1.

2. The Conjunction Principle

The conjunction principle states that there is no α, where $1 > \alpha > 0$, such that $Pr(A|C) \geq \alpha$ and $Pr(B|C) \geq \alpha$ necessitate that $Pr(`AandB'|C) \geq \alpha$. Given the definitions and explanations in the previous appendix, the following figure verifies the above claim when α is 0.5. And this results generalizes to any value smaller than 1. The conditional probabilities that are represented in the figure shows that Pr ('A and B' | C) could be as low as zero when Pr (A | C) and Pr (B | C) are both 0.5.

Pr (A | C) = 50 / (50 + 0 + 50) = 50 /100 = 0.5

Pr (B | C) = 50 / (50 + 0 + 50) = 9 /100 = 0.5

Pr ('A and B' | C) = 0 / (50 + 0 + 50) = 0 /100 = 0.0

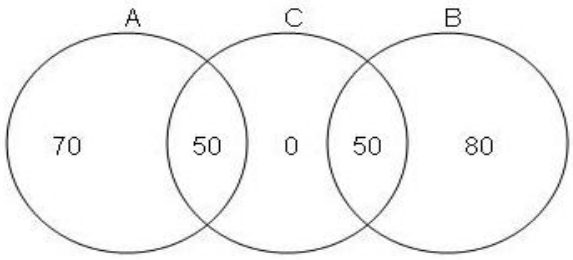

FIGURE 3. Conjunction Principle

The conjunction principle holds when Pr (A | C) and Pr (B | C) are set to 1. Figure 2 in Appendix 1 depicts such a case.

3. Dretske's Entropy Calculations

Dretske in his 22^{nd} footnote for Chapter 2 of Knowledge and the Flow of Information assumes three stations A, B and C. For each station, there are eight equally likely possibilities. Sine the amount of information that an event carries is equal the logarithm of the inverse of its probability, if an event occurs at A, B or C it carries 3 bits of information. Using logarithm base 2,

$I(c_2) = \log \frac{1}{Pr(c_2)} = \log \frac{1}{\frac{1}{8}} = \log 8 = \log 2^3 = 3$

c_2 occurs at C and the information flows from C to B. Conditional probabilities between B and C are such that the occurrence of b_2 at B raises the probability that c_2 occurred to 0.9 and lowers the probability of others to 0.014 which is equal to $\frac{1-0.9}{7}$. This implies that there is equivocation between B and C. The formula for calculating the amount of equivocation is the following.

$E(b_2) = -\sum_(i = 1)^n Pr(c_i|b_2) \times \log Pr(c_i|b_2)$

This equation gives us 0.22 bits equivocation. Hence the amount of information transferred to B from C is 3 - 0.22 = 2.78 bits. Assuming that B and A are identical stations, that is to say the occurrence of a_2 in A increases the probability that b_2 occurred to 0.9 an lowers the others to 0.014, the equivocation between A and C is calculated by using

the above equation. This equivocation, i.e. E(a_2), is 1.3 bits. Hence, a_2 carries 1.7 bits of information about what happened at C.

Dretske says that this is paradoxical, since b_2 tells us what happened at C, and a_2 tells us what happened at B, but a_2 cannot tell us what happened at C. The information loss between A and C (1.3 bits) is significantly higher than the information losses between A and B, and between B and C (both 0.22).

4. The Ordinal Ranking Approach & The Conjunction Principle

For the sake of simplicity, assume that the signal r carries two pieces of information: Red and Sphere. In other words, a red and spherical object sends a signal to an information processing organism. The organism receives the signal r, and thereby the information that r is red and that r is spherical separately. The question is whether the organism has the information that r is both red and spherical. Moreover, assume that the organism's color categories include only Red, Green and Yellow, and its shape categories include Sphere, Pyramid and Cylinder. These would form the basis for the sets of competing hypotheses. In the ordinal ranking approach, in order for the signal r to carry the information that it is red, the category Red should have the highest conditional probability over all the other competing categories. Let's assign the following conditional probabilities to ensure that the signal r carries the information that is red which is a part of the required assumption of the Conjunction Principle. These numbers are also stated diagrammatically in Figure 4.

Pr ('Red' | r) = 45 / (45 + 30 + 25) = 45 /100 = 0.45

Pr ('Green' | r) = 30 / (45 + 30 + 25) = 30 /100 = 0.30

Pr ('Yellow' | r) = 25 / (45 + 30 + 25) = 25 /100 = 0.25

The signal r should also carry the information that it is Sphere. That is to say, the conditional probability of Sphere given r should be higher than the competing categories. The following assigned numbers satisfy this assumption, and Figure 5 visually shows that distribution.

Pr ('Sphere' | r) = 40 / (40 + 35 + 25) = 40 /100 = 0.40

Pr ('Pyramid' | r) = 35 / (40 + 35 + 25) = 35 /100 = 0.35

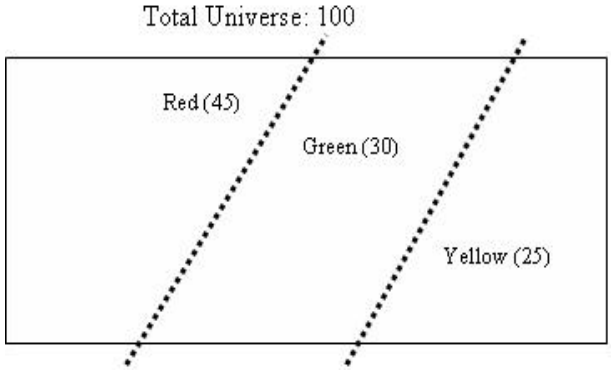

FIGURE 4. First Property

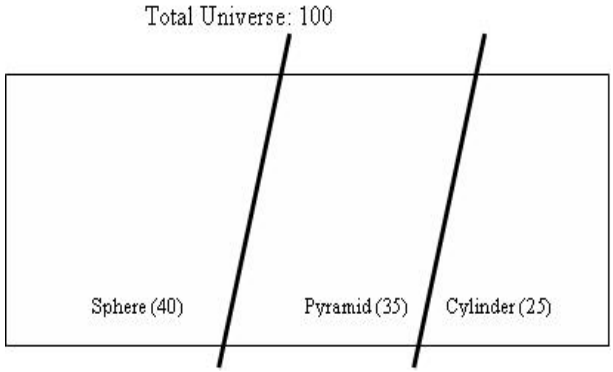

FIGURE 5. Second Property

Pr ('Cylinder' | r) = 25 / (40 + 35 + 25) = 25 /100 = 0.25

If there is no independency between the color dimension and the shape dimension, then it is possible for r not to carry the information that Red and Sphere, because of the possible overlap between the regions in Figure 4 and Figure 5. Hence, it is possible for r not to carry the information that Red and Sphere while carrying the information that Red and the information that Sphere separately. This is an unacceptable consequence and the basis of Dretske's argument from the Conjunction Principle for assigning unity to conditional probabilities. However, if there is an independency constraint, which we have as a result of the notion of minimal inconsistency, then the overlap is avoided. As a result, the regions in

Figures 4 and 5 become additive. The total universe becomes 200, and the region for Red and Sphere becomes 85 which is greater than the regions for all other pairs.

Total Universe = 45 + 30 + 25 + 40 + 35 + 25 = 200

Pr ('Red and Sphere' | r) = (45 + 40) / 200 = 85 /200 = 0.425

Pr ('Red and Pyramid' | r) = (45 + 35) / 200 = 80 /200 = 0.4

Pr ('Red and Cylinder' | r) = (45 + 25) / 200 = 70 /200 = 0.35

Pr ('Green and Sphere' | r) = (30 + 40) / 200 = 70 /200 = 0.35

Pr ('Green and Pyramid' | r) = (30 + 35) / 200 = 65 /200 = 0.325

Pr ('Green and Cylinder' | r) = (30 + 25) / 200 = 55 /200 = 0.275

Pr ('Yellow and Sphere' | r) = (25 + 40) / 200 = 65 /200 = 0.325

Pr ('Yellow and Pyramid' | r) = (25 + 35) / 200 = 60 /200 = 0.3

Pr ('Yellow and Cylinder' | r) = (25 + 25) / 200 = 50 /200 = 0.25

Bibliography

Alhazen, A. (1989). In A. Sabra, ed., *The Optics of Ibn al-Haytham*, Warburg Institute.

Bar-Hillel, Y. (1955). An examination of information theory. *Philosophy of Science*, **22**, 86–105.

Brooks, D.R. & Wiley, E.O. (1988). *Evolution as Entropy: Toward a Unified Theory of Biology*. University of Chicago, 2nd edn.

Carnap, R. & Jeffrey, R. (1971). *Studies in Inductive logic and Probability*, vol. 2. University of California Press.

Chalmers, D.J. (1996). *The Concious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Clark, A. (1993). Mice, shrews and misrepresentation. *Journal of Philosophy*, **90**, 290–310.

Clark, A. (2000). *Mindware : An Introduction to the Philosophy of Cognitive Science*. Oxford University Press.

Cohen, J. & Meskin, A. (2006). An objective counterfactual theory of information. *Australasian Journal of Philosophy*.

Cover, T.M. (1991). *Elements of Information Theory*. Wiley Interscience Publications.

Crumley, J.S. (1999). *Problems in Mind: Readings in Contemporary Philosophy of Mind*. McGraw-Hill Humanities.

Cummins, R. (1991). *Meaning and Mental Representation*. MIT Press.

Cummins, R. (1996). *Representations, Targets and Attitudes*. MIT.

Cummins, R. & Poirier, P. (2004). Representation and indication. In P. Staines & P. Slezak, eds., *Representation in Mind*, Elsevier.

Denkel, A. (1995). Meaning, communication and behavior. In *Reality and Object*, Bogazici University Publications.

Denkel, A. (1996). *Object and Property*. Cambridge University Press.

Dennett, D.C. (1969). *Consciousness Explained*. Back Bay Books.

Descartes, R. (1984-85a). Meditations on first philosophy. In J. Cottingham, R. Stoothoff & D. Murdoch, eds., *Philosophical writings of Descartes*, Cambridge University Press.

Descartes, R. (1984-85b). Optics (selections). In J. Cottingham, R. Stoothoff & D. Murdoch, eds., *Philosophical writings of Descartes*, Cambridge University Press.

Dretske, F. (1983). Author's response. *Behavorial and Brain Sciences*, **6**.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. The MIT Press.

Dretske, F. (1994). Misrepresentation. In S. Stich & T. Warfield, eds., *Mental Representation: A Reader*, Blackwell.

Dretske, F.I. (1981). *Knowledge and the Flow of Information*. The MIT Press.

Eliasmith, C. (2000). *How Neurons Mean: A Neurocomputational Theory of Representational Content*. Ph.D. thesis, Washington University in St. Louis.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*, Elsevier Science B.V.

Fano, R.M. (1961). *Transmission of Information: a Statistical Theory of Communications*. MIT Press.

Fitzhugh, R. (1958). A statistical analyzer for optic nerve messages. *Journal of General Physiology*, **41**.

Fodor, J. (1975). *Language Thought*. New York: Crowell.

Fodor, J. (1983). *Modularity of mind: An essay on faculty psychology*. MIT Press.

Fodor, J. (1984). Semantics: Wisconsin style. *Synthese*, **59**.

Fodor, J. (1994). Fodor's guide to mental representation. In S.P. Stich & T.A. Warfield, eds., *Mental Representation: A Reader*, Blackwell Publishers.

Fodor, J. & Lepore, E. (1992). *Holism: A Shopper's Guide*. Blackwell.

Fodor, J.A. (1992). *A Theory of Content and Other Essays (Bradford Books)*. The MIT Press.

Gardner, H. (1984). *The Mind's New Science*. Basic Books.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Hougton Mifflin.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Harvard University Press.

Grandy, R.E. (1987). Information-based epistemology, ecological epistemology and epistemology naturalized. *Synthese*, **70**, 191–203.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hajek, A. (1997). Mises redux: Fifteen arguments against finite frequentism. *Erkenntnis*, **45**, 209–227.

Harms, W.F. (1998). The use of information theory in epistemology. *Philosophy of Science*, **65**, 472–501.

Hatfield, G. (2002). Perception as unconcious inference. In D. Heyer & R. Mausfeld, eds., *Perception and the Physical World: Psychological and Philosophical Issue in Perception*, Wiley and Sons Ltd.

Helmholtz, H.L. (1971). In R. Kahl, ed., *Selected writings of Hermann von Helmholtz*, Wesleyan University Press.

Helmholtz, H.L. (1995). In D. Cahan, ed., *Science and culture: Popular and philosophical essays by Hermann von Helmholtz*, University of Chicago Press.

Hintikka, J. (1970). On semantic information. In J. Hintikka & P. Suppes, eds., *Information and Inference*, Dordrecht.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**, 106–154.

Kant, I. (2004). *Metaphysical Foundations of Natural Science*. Cambridge Texts in the History of Philosophy, Cambridge University Press.

Kyburg, H.E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.

Kyburg, H.E. (1983). Knowledge and the absolute. *Behavorial and Brain Sciences*, **6**.

Lehrer, K. (1970). Justification, explanation and induction. In M. Swain, ed., *Induction, Acceptance, and Rational Belief*, Dordrecht: Reidel.

Lettvin, J., McCulloch, W. & Pitts, W. (1940). What the frog's eye tells the frog's brain? *Proc. IRE*, **47**.

Loewer, B. (1983). Information and belief. *Behavorial and Brain Sciences*, **6**.

Loewer, B. (1987). From information to intentionality. *Synthese*, **70**.

Mason, L. (1972). *Wolf Children and the Problem of Human Nature*. Monthly Review Pr.

McLeod, P., Plunkett, K. & Rolls, E.T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford University Press.

Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, **86**, 288–302.

Nagel, T. (1989). *The View from Nowhere*. Oxford University Press.

Olshausen, B.A. & Field, D.J. (2005). How close are we to understanding v1? *Neural Computation*, **17**, 1665–1699.

Ptolemy, C. (1996). In A. Smith, ed., *Ptolemys theory of visual perception: An English translation of the Optics*, American Philosophical Society.

Reichenbach, H. (1949). *The theory of probability, an inquiry into the logical and mathematical foundations of the calculus of probability*. University of California Press, Berkeley.

Rieke, F., Warland, D., deRuytervanSteveninck, R. & Bialek, W. (1999). *Spikes: Exploring the Neural Code (Computational Neuroscience)*. The MIT Press.

Rock, I. (1975). *Introduction of Perception*. Macmillan Publishing Co.

Rock, I. (1983). *Logic of Perception*. MIT Press.

Rowlands, M. (1999). Teleosemantics. *Online A Field Guide to Philosophy of Mind*.

Salmon, W. (1966). *The Foundations of Scientific Inference*. University of Pittsburgh Press.

Scarantino, A. (2005). Did dretske learn the right lesson from shannon's theory of information?, manuscript.

Shannon, C. & Weaver, W. (1980). *The Mathematical Theory of Communication*. University of Illinois Press, 8th edn.

Stampe, D. (1977). Towards a causal theory of linguistic representation. In P. French, T. Euhling & H. Wettstein, eds., *Midwest Studies in Philosophy 2*, The University of Minnesota Press.

Uexkull, J.V. (1926). *Theoretical Biology*. Kegan & Trench & Trubner & Co.

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Language*, **16**, 311–334.

Wheeler, J.A. (1994). *It from Bit*, 295–312. American University of Physics Press.

Wiener, N. (1961). *Cybernetics, or Control and Communication in the Animal and the Machine*. MIT Press, 2nd edn.

DEPARTMENT OF PHILOSOPHY, SYCAMORE HALL 026, 1033 E. THIRD ST.,
BLOOMINGTON, IN 47405-7005, USA
PHONE (812) 360 9247 E-MAIL hdemir@indiana.edu
WEBPAGE www.benhilmi.com
FAX (812) 855 3777

HILMI M. DEMIR

AREAS OF SPECIALIZATION & COMPETENCE

- **AOS**: Philosophy of Mind, Philosophy of Information, Philosophy of Psychology and Cognitive Science
- **AOC**: Metaphysics, Logic, Philosophy of Science, Ancient Philosophy

PROFESSIONAL EXPERIENCE

- **Assistant Professor** — **Philosophy Depart.** — **California State Univ., San Bernardino**
  Starts in Fall 2006 — *Tenure – Track*
- **Assistant Instructor** — **Indiana Univ., Bloomington**
  Fall 2001 – Spring 2006 — *Philosophy & Cognitive Science Departments*

EDUCATION

- Ph.D. in Philosophy — August 2006 — Indiana University, Bloomington
- Ph.D. in Cognitive Science — August 2006 — Indiana University, Bloomington
  - Dissertation: ***Error Comes with Imagination: A Probabilistic Theory of Mental Content***
  - Philosophy Chair: Prof. Frederick Schmitt
  - Cognitive Science Chair: Prof. Colin Allen
- M.A. in Philosophy — June 1999 — Boğaziçi University, Istanbul, Turkey
  - Thesis: ***A Philosophical Examination of the Concept of Memory: Aristotle & Hume***
  - Chair: Prof. Gürol Irzık
- B.A. in Philosophy — Boğaziçi University, Istanbul, Turkey
  - B.A. in philosophy. Ranked 1st among all graduates. Honor Degree

PUBLICATIONS

- Allwein G., Demir H., Pike L. "Logic Classes for Boolean Monoids and an Interpretation in CMOS Circuits", ***Journal of Logic, Language and Information*** Vol 13, No: 3, December 2004, pp. 241-266, Kluver Academic Publishers. **(Refereed Paper)**

- Sheya A., Demir H., Hanania R. "Consistent Argument Predicate Binding is Important for Predicate-Predicate Linking", ***Proceedings of the 26th Annual Cognitive Science Society***, August 2004, Lawrence Erlbaum Associates. **(Refereed Paper)**

CONFERENCE PRESENTATIONS

- "Cognitive Science as an Independent Discipline: A SWOT Analysis Attempt" **Cognitive Science Curriculum Workshop**, Indiana University, June 2006 (**Invited Speaker**)

- "Inverse Conditional Probabilities: Good or Evil?", ***Neurophilosophy: The State of the Art***, McDonnell Project, California Institute of Technology, June 2005 (**Refereed poster presentation**)

- "(Mis)Representation and Propositional Content", ***Cognitive Science Group Lecture Series***, California State University Fresno, April 2005 (**Invited speaker**)

- "Error Comes With Imagination: A Probabilistic Theory of Mental Content", ***The Midsouth Philosophy Conference***, University of Memphis, February 2005 (**Refereed paper**)

    - **Commentator on** Gregory Johnson's "Comments on Affect Programs", ***The Midsouth Philosophy Conference***, February 2005

- "A Worldly Wittgenstein: Situation Semantics as a Model for Tractarian Awe" ***The History of Analytic Philosophy Conference***, University of Iowa, 2002 (**Refereed paper**)

- "Influential Ideas – Weak Arguments: Frege against Psychologism", ***The Foundations of Mathematics Conference,*** the Middle East Technical University, Turkey, 2000 (**Invited speaker**)

AWARDS – FELLOWSHIPS

- **Chancellor's Fellowship:** 2000 – 2004 — Indiana Univ., Bloomington (IUB)
- **Outstanding Instructor Award:** 2004 — Philosophy Dept, IUB
- **Summer Research Awards:** 2002 through 2006 — Cognitive Science Dept, IUB
- **Outstanding Service Award:** 2003 — Philosophy Dept, IUB

TEACHING EXPERIENCE

- **Instructor:** with the full responsibility of designing the course, lecturing, grading and in some cases supervising teaching assistants.
  - Philosophical Foundations of Cognitive & Information Science
    - Fall 2005 & Fall 2004 IUB (in both of these, supervision of a teaching assistant)
  - Introduction to Symbolic Logic
    - Summer 2004 IUB
  - Elementary Logic
    - Fall 2003 IUB (supervision of a teaching assistant)
  - Critical Thinking and Reasoning
    - Summer 2005 IUPUI & Fall 2003 IUB
  - Introduction to Philosophy
    - Fall 1999, University for Freedom, Istanbul, Turkey
  - Philosophy, Science and Art
    - Spring 2000, University for Freedom, Istanbul, Turkey

- **Teaching Assistant:** with the responsibility of leading the discussion sections, grading and significant input in designing the course.
  - Math & Logic for Cognitive Scientists (Graduate level, worked with Prof. Larry Moss)
    - Spring 2006 IUB
  - Math & Logic for Cognitive Scientists (Graduate level, worked with Prof. Larry Moss)
  - Logical Theory I (Graduate level, worked with Joan Weiner as the grader)
    - Fall 2004 IUB
  - Elementary Logic (worked with David McCarty)
    - Spring 2005 IUB
  - Introduction to Ethics (worked with Marcia Baron)
    - Spring 2003 & Fall 2002 IUB
  - Philosophical Foundations of Cognitive Science & Information Science (worked with Brian Bowdle)
    - Spring 2002 & Fall 2001 IUB
  - Philosophy of Mind (worked with Güven Güzeldere)
    - Summer 1999, Boğaziçi Univ., Istanbul, Turkey
  - Introduction to Logic I & II
    - Fall 1998, Spring 1999, Fall 1999, Spring 2000, Boğaziçi Univ., Istanbul, Turkey
    - Worked with Berna Kılınç, Ilhan Inan and Karanfil Soyhun.

REFEREE WORK

- ***Journal of Mathematical Psychology,*** 2003, ed. Jim Townsend
- ***Connection Science,*** 2003 Assoc. Ed. Andy Clark

SERVICE

- Philosophy of Cog. Sci. Senior Hire Search Committee — 2004 – 2005, Philosophy, IUB
- University Graduate Student Organization Representative — 2003 – 2004, Philosophy, IUB
- Philosophy Graduate Admissions Committee — 2002 – 2003, Philosophy, IUB
- International Student Orientation Session Presenter — 2003, IUB
- Philosophy of Cog. Sci. Junior Hire Search Committee — 2002 – 2003, Philosophy, IUB

PROFESSIONAL AFFILIATIONS

- American Philosophical Association
- Cognitive Science Society
- Society for Philosophy & Psychology

REFERENCES

- Fredrick Schmitt — IUB
  Professor of Philosophy
  E-mail : fschmitt@indiana.edu, Phone : + 1 – 812 – 855 – 3296
- Richard Shiffrin — IUB
  Distinguished Professor
  Luther Dana Waterman Professor of Psychology and Director of Cognitive Science
  E-mail : shiffrin@indiana.edu
- Gerard Allwein — NAVY Research Lab, Washington D.C.
  Senior Researcher
  allwein@itd.nrl.navy.mil
- Colin Allen — IUB
  Professor of History and Philosophy of Science, Cognitive Science Department and Center for the Integrative Study of Animal Behavior
  E-mail : colallen@indiana.edu, Phone : + 1 – 812 – 855 – 8916
- Tim O'Connor — Philosophy, IUB
  Professor of Philosophy
  E-mail: toconnor@indiana.edu
- David McCarty — Philosophy, IUB
  Professor of Philosophy
  E-mail: dmccarty@indiana.edu
- Jonathan Weinberg — Philosophy and Cognitive Science, IUB
  Assistant Professor of Philosophy
  E-mail : jmweinbe@indiana.edu